Identity by descent analysis with applications to epilepsy studies and *Plasmodium* causing human malaria

A thesis submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

by

Lyndal Jane Henden

Department of Medical Biology

The University of Melbourne

Population Health and Immunity Division

The Walter and Eliza Hall Institute of Medical Research

February 2017

Abstract

Relatedness mapping is concerned with identifying genomic regions that have been inherited from a common ancestor. Such regions are said to be identical by descent (IBD) and detecting IBD has proven useful in many applications including disease mapping, discovery of familial relatedness and determining loci under selection. Relatedness mapping is typically performed on humans. As such, methodologies are widely available for diploid genomes. This readily allows for analysis of autosomal chromosomes; however, algorithms are generally not applicable to the X chromosome. This is because females have two copies of the X chromosome while males have one copy of the X chromosome, requiring a more complicated model to account for the difference in chromosomal numbers between males and females. As a result, the X chromosome is generally excluded from analysis. This is unfortunate as an abundance of disorders, such as intellectual disability, epilepsy and autism, would greatly benefit from X chromosome IBD analysis.

This thesis describes the first probabilistic methodology for identifying pairwise IBD that is applicable to both the X chromosome and autosomes. Genotype data, extracted from either SNP arrays or next generation sequencing platforms, is used to infer IBD and issues surrounding genotyping errors, missing data and linkage disequilibrium are accounted for. Statistical and bioinformatics analyses are carried out to demonstrate the performance of the methodology and an analysis of a small cohort of individuals with a rare form of epilepsy is successfully performed.

The lack of methodologies for haploid chromosomes has implications that extend beyond the context of the X chromosome. In particular, microorganisms with haploid genomes, such as the malaria causing parasite, *Plasmodium*, and bacterium *Staphylococcus aureus*, are unable to be analysed for relatedness. Such analyses would be invaluable for the study of these, and other, diseases with the recent emergence of antimicrobial drug resistance, whereby a number of microorganisms have become resistant to first-line and/or

last-resort antimicrobial treatments. IBD can be used to identify loci under selection that are associated with antimicrobial resistance as well as monitor the genetic diversity in populations to track disease control efforts.

One of the main difficulties with analysing genomic data of microorganisms, collected from a human infection, is the occurrence of multiple genetically-distinct strains within an infection. The genomic data can no longer be treated as though it is haploid and requires special treatment. Like the human X chromosome, these samples are commonly excluded from analysis, which can greatly reduce the power of studies in regions where multiple infections are common.

As such, this thesis additionally describes the necessary extensions for the X chromosome IBD methodology to be applicable to non-human haploid organisms, where multiple infections may be present. The algorithm successfully identifies antimalarial drug resistance loci under positive selection in a global dataset of the deadliest species of malaria, *P. falciparum*, which has recently become resistant to the first-line treatment, prompting a global health crisis. This thesis also demonstrates the valuable insights that can be gained from IBD analysis of malaria, which are applicable to other infectious diseases.

Declaration

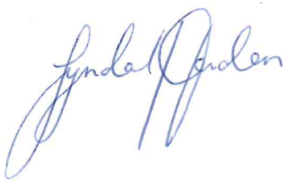
This is to certify that

i. the thesis comprises only my original work towards the degree of Doctor of Philosophy except where indicated in the Preface,

ii. due acknowledgement has been made in the text to all other material used,

iii. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed,



Preface

Several chapters of work in this thesis were carried out in collaboration with others. The methodology in Chapter 2 has been published in the following:

Henden, L., Wakeham, D. and Bahlo, M. (2016). XIBD: software for inferring pairwise identity by descent on the X chromosome. *Bioinfomatics*. 32(15): 2389-2391

Professor Melanie Bahlo conceived ideas and provided guidance while Mr David Wakeham performed some initial mathematics. I finalized the mathematics, developed the software and performed all statistical and bioinformatics analyses.

The results from Chapter 4 have been published in the following:

Henden, L. *et al.* (2016). Identity by descent fine mapping of familial adult myoclonus epilepsy (FAME) to 2p11.2-2q11.2. *Hum Genet*. 135(10):1117-1125

The primary collaborators involved in this work were Dr Mark Corbett and Dr Saskia Freytag. Dr Mark Corbett provided us with array genotyping data and made decisions on behalf of the FAME consortium, constituting authors numbered 3 to 28. Participant recruitment, clinical and laboratory work were carried out by the members of the FAME consortium. Dr Saskia Freytag performed gene prioritization and I conducted all other bioinformatics inferences, constituting 90% of the work.

The results from Chapter 6 are available as a preprint in the following:

Henden, L., Lee, S., Mueller, I., Barry, A. and Bahlo, M. (2016). Detecting selection signals In Plasmodium falciparum using identity-by-descent analysis. *BioRxiv*. doi: https://doi.org/10.1101/088039

I collaborated primarily with Mr Stuart Lee. The publicly available MalariaGEN Pf3k data was use in this analysis as well as the MalariaGEN sequence data from Papua New Guinea. Mr Lee performed all pre-analysis data processing procedures and implemented selection tools. Remaining authors provided valuable insights and discussions into the results. I developed the software used in this work and performed all statistical and bioinformatics analyses.

I wrote the first drafts of the work presented in Chapters 2, 4 and 6.

Additional work was performed as part of this PhD that is not included in this thesis. This includes identity by descent analyses in the following publications;

Shaw, M., Yap, T Y., **Henden, L.**, Bahlo, M. *et al.* (2015). Identical by descent L1CAM mutation in two apparently unrelated families with intellectual disability without L1 syndrome. *Eur J Med Genet.* 58(6):364-368

Skopkova, M., Hennig, F. *et al.* (2017). EIF2S3 mutations associated with severe X-linked intellectual disability syndrome MEHMO. *Hum Mutat.* doi: 10.1002/humu.23170

I performed less than 50% of the work in each of these publications.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Melanie Bahlo, for allowing me to do my PhD with her. Melanie, thank you for your patience and guidance throughout this journey. I have gained a wealth of knowledge under your supervision and am continually inspired by your passion and enthusiasm for your work. This has been the most rewarding experience of my life to date and I am extremely grateful for the opportunities you have provided me with and the time you have dedicated to my growth as a research scientist. I now have a greater sense of confidence and pride in myself, and for that I cannot thank you enough.

I would also like to thank my co-supervisor, Professor Terry Speed, without whom I would not have met Melanie. Terry, thank you for your encouragement, positivity and guidance throughout this endeavor. To the past and present members of the Bahlo lab, thank you for your help and encouragement during this journey. To Peter and Stuart, I am extremely grateful for the time you have dedicated to my research. Your knowledge, encouragement and assistance has been invaluable. I am also very grateful for the support from the WEHI Bioinformatics Division and Population Health and Immunity Division. Your assistance and constructive ideas have been invaluable to my research. To my advisory committee, thank you for believing in me and making time in your busy schedules to attend my presentations.

I thank the Australian Postgraduate Award and the John and Patricia Farrant foundation, whose funding has enabled me to fully devote my time to this research. I would also like to thank WEHI for their generosity as well as Edith Moffat for supporting my travels overseas to attend international conferences and explore research institutes.

My time in Melbourne has been truly amazing, and this is largely because of the incredible friendships I have made. To Maddy and Nellie, our fortnightly brunch dates could not be regular enough. I am extremely grateful to have two incredibly kind and

understanding friends who balance my lifestyle with food, laughter, gossip and wine. To the crew who keeps me regularly fed on a Wednesday night, you are some of the most wonderful and intelligent people I have ever met. It is a privilege to call you my friends.

To my friends in New Zealand who have made my trips home that little bit more special. Marina and Katie, your loving friendship has become more valuable to me as the years pass by. Thank you for the always making the effort to catch-up at busy times of the year, I will never get tired of drinking tea and coffee in the old stomping ground with you both. To my dear friend Alison, thank you for reassuring me of my abilities over our many phone calls. Your cheerful presence and positive attitude was a great comfort, particularly in the more stressful times of this journey.

To my superstar partner, Michael. Thank you for celebrating all of my small and large victories with ice cream and froyo. I am incredibly lucky to have such a loving and thoughtful person to share this experience with.

Finally, I thank my wonderful family for their unparalleled love and continuous support throughout this endeavour. To mum and dad, you have been a constant source of encouragement over the many years of study. I am extremely grateful for your knowledge, wisdom, and guidance with all of life's smallest and largest ventures. From expert cooking tips for the perfect chocolate eclairs to crossing the Tasman Sea; I could not ask for more loving and devoted parents. To my sister and best friend, you have been the most amazing support of all. Since the emotional day that I left, you have encouraged me every step of the way. I am forever grateful to have such a kind, generous and loving sister who makes me a better, more fun and sometimes sillier person. I have found no greater comfort than in sharing all of life's experiences with you.

Contents

List of Figures

List of Tables

List of Abbreviations

bp	Base pair
cM	centiMorgan
DNA	Deoxyribonucleic acid
EHH	Extended haplotype homozygosity
FAME	Familial adult myoclonus epilepsy
GPS	Grey platelet syndrome
GWAS	Genome-wide association studies
HLA	Human leukocyte antigen
HMM	Hidden Markov Model
IBD	Identity by descent
IBS	Identity by state
LD	Linkage disequilibrium
LE	Linkage equilibrium
LLR	Log-likelihood ratio
M	Morgan
MAF	Minor allele frequency
Mb	Mega base pairs (million base pairs)
MOI	Multiplicity of infection
MQ	Mapping quality
NGS	Next generation sequencing
PNG	Papua New Guinea
QD	Quality of depth
SNP	Single nucleotide polymorphism

SOR	Strand odds ratio
TSI	Toscani in Italy
WEHI	The Walter and Eliza Hall Institute of Medical Research
WGS	Whole-genome sequencing

Chapter 1

Introduction

This chapter provides an overview of the terminology required for understanding this thesis. It introduces basic biology, identity by descent analysis and the current methodologies that can infer genomic regions shared identical by descent.

1.1 A recombining genome

The genome contains all the necessary genetic information for the development and functioning of a living organism[1]. This information is encoded as deoxyribonucleic acid (DNA) and is composed of four chemical bases; adenine (A), guanine (G), cytosine (C) and thymine (T). Each base couples with its complement (A with T and C with G) to form a *base pair*, which combines with a sugar and phosphate molecule to produce a *nucleotide*. Many nucleotides are linked together in a sequence-like structure to form the backbone of DNA. DNA is located within the nucleus of eukaryotic cells and is packaged into smaller units called *chromosomes*. Each chromosome contains a number of short segments of DNA, known as *genes*, that encode the physical traits of a species. Slight variations within a gene can result in different observable traits, introducing genetic diversity into a population. The variations of a gene or genomic location, called *locus* (plural: *loci*), are referred to as *alleles* and allow species to adapt to changing environments.

1.1.1 Ploidy, meiosis and recombination

The number of sets of chromosomes within a cell is referred to as the cell *ploidy*. A *haploid* cell contains one complete set of chromosomes and by extension, an organism is considered haploid if all of its cells are haploid. Such is the case for male bees[2] and

ants[3], as well as microorganisms including the bacterium *Mycobacterium tuberculosis*[4] and the malaria causing parasite *Plasmodium*[5]. In contrast, *diploid* cells contain two sets of chromosomes, where chromosomes from each set form homologous pairs based on the similarity of chromosome lengths and gene locations. In diploid organisms, all cells need not necessarily be diploid for an organism to be considered diploid. For example, nearly all mammals are diploid, although a small proportion of cells are usually haploid[6].

Common types of haploid cells are gametes (egg or sperm cells) and these are produced during sexual reproduction in a process called *meiosis* (Figure 1.1)[1]. Initially, two gametes, one from a male and one from a female, fuse together to create a diploid embryo. Chromosomes from each set form homologous pairs based on the similarity of the chromosome lengths and gene locations. These homologous chromosomes then duplicate such that each chromosome consists of two identical strands, called sister chromatids, and each homologous pair consists of four stands. The homologous chromosomes then undergo synapsis whereby each pair connects to one another and the sister chromatids begin to crossover. Here, part of a chromatid on one chromosome breaks off and switches position with the matching portion of a chromatid on the homologous chromosome. This exchange of genetic material via crossing over is referred to as *recombination* and results in offspring with different combinations of genes to their parents. Following recombination, the homologous chromosome pairs randomly align in the cell center where they are then pulled in opposite directions. Cell division occurs and two haploid daughter cells are formed, each with one complete set of chromosomes, where the sister chromatids remain attached. In a similar manner, the sister chromatids within each daughter cell randomly align in the cell's center and are separated to opposite sides of the cell where cell division follows. The two former daughter cells each divide into two additional haploid cells, generating a total of four gametes.

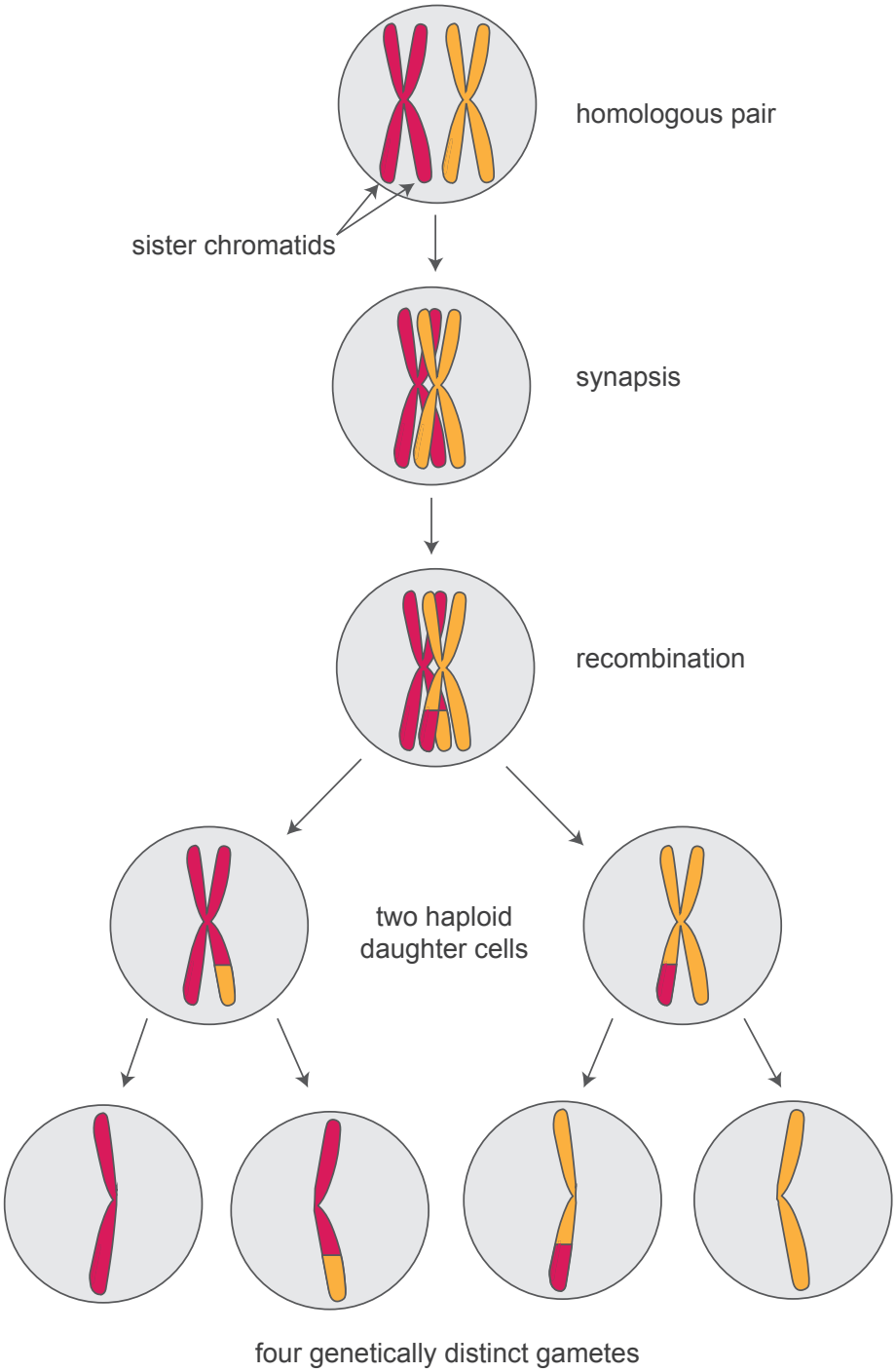


Figure 1.1: The process of meiosis in diploid cells. A homologous pair of chromosomes undergo synapsis whereby the sister chromatids of homologous pairs cross over and exchange genetic material. The recombined homologous chromosomes divide creating haploid daughter cells, which divide again to produce four genetically distinct haploid gametes. This figure was adapted from Griffiths et al.[1] and drawn using Adobe Illustrator.

Meiosis is of great importance in diploid organisms as it is a key source of genetic diversity that allows for a wider variety of genetic traits to be selected within a population. This is achieved through the fundamental process of recombination. The genomic

locations at which recombination occurs are called *crossover points* and there is thought to be at least one crossover point per chromosome pair during meiosis[7]. While crossover events can occur anywhere along a chromosome, they are generally suppressed near the chromosome center (*centromere*) and amplified towards the chromosome ends (*telomeres*). Additionally, loci in close proximity to an existing crossover point are less likely to experience another crossover in a single meiosis event. Given this relationship, the expected number of crossovers between two loci on a single chromosome can be used as a measure of genetic distance, called the *genetic map distance*[7]. The unit of genetic map distance is *Morgan* (M), or more commonly used *centiMorgan* (cM), where 1 M = 100 cM. Formally, 1 cM is the probability of observing a recombination event between two loci during a single meiosis[7]. In humans, 1 M $\approx$ 1,000,000 bp (or 1 Mb), on average (Figure 1.2). Other species show a different relationship.

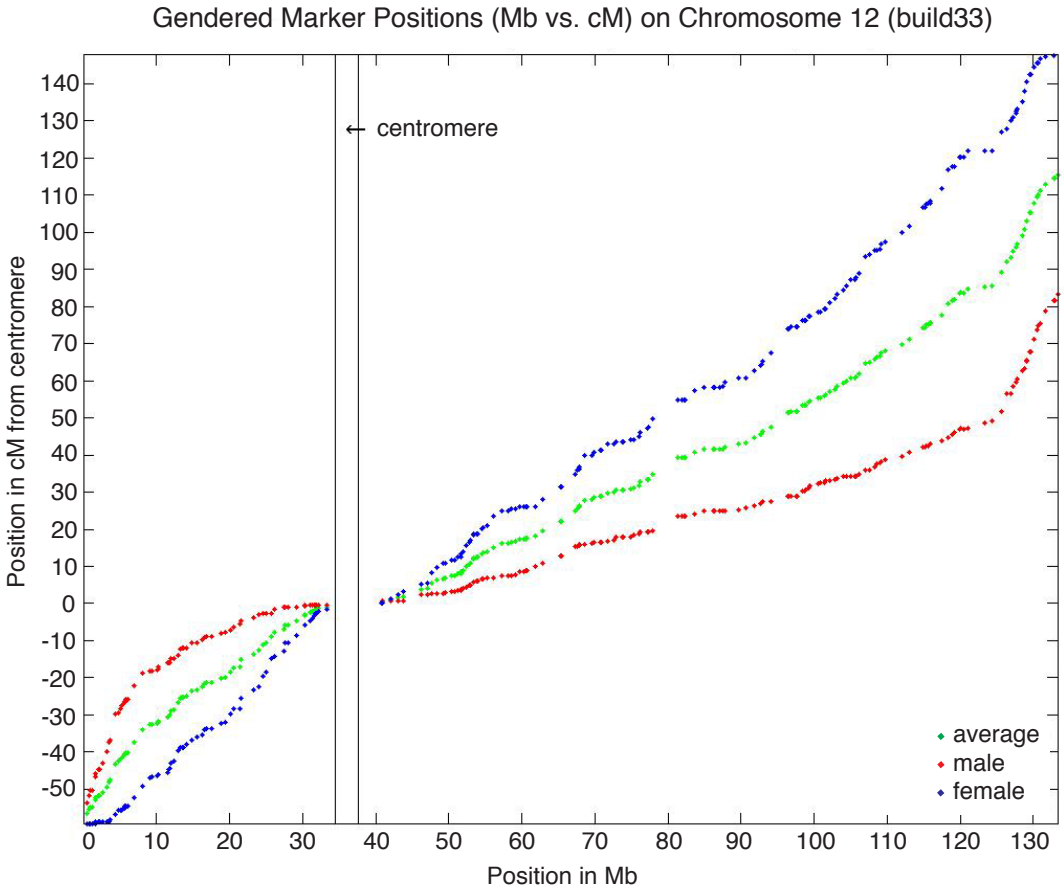


Figure 1.2: The relationship between the genetic map distance (cM) and the physical distance (bp) on chromosome 12 for males, females and averaged by sex. This figured was sourced from Lander et al.[8].

1.1.2 Single nucleotide polymorphisms

In addition to recombination, there are a number of other sources of genetic variation that can be more broadly classified as mutations. A mutation is an alteration in the DNA sequence that includes deleting, inserting, substituting, or rearranging a section of DNA, among other things[1]. The most common type of mutation is a *single nucleotide polymorphism*, or SNP, which is simply a change in a single base pair[9]. For the purpose of this thesis, only SNPs will be discussed. Additionally, only SNPs with two alleles present, the *reference* (*wild type*) allele or *alternative* allele, will be covered. Such SNPs are said to be *biallelic* and constitute the majority of SNPs in the human genome.

Most SNPs are passed down from one generation to the next and only a small number are produced during meiosis. Many of these SNPs lead to no observable difference in a species, while some SNPs account for differences in appearances and others affect how a species develops diseases[1]. Given the heritability of SNPs, it follows that SNPs tend to be conserved within a population, however SNPs can vary considerably between populations[10]. For example, a locus that is polymorphic in one population may be monomorphic (does not vary) in another population. Alternatively, a SNP may be present in two populations with the wild type allele appearing at different frequencies.

A sequence of SNPs (not necessarily adjacent or uniformly spaced) on a chromosome that have been inherited from a single parent form a *haplotype*. It follows that a haplotype extracted from a haploid chromosome is phased, while phasing algorithms are typically required to construct the haplotypes for diploid chromosomes. The unphased combinations of alleles at SNPs are instead referred to as *genotypes*. The genotype of a biallelic SNP on a diploid chromosome can be classified as homozygous reference, heterozygous or homozygous alternative, referring to zero, one or two copies of the alternative allele, respectively. In contrast a haploid chromosome can have either zero or one copy of the alternative allele at a SNP. Additionally, the genotype of a haploid chromosome is simply the haplotype.

1.1.3 The human genome

The human genome is organized into 23 pairs of chromosomes encompassing more than 3 billion base pairs and ranging in size from 48 Mb to 250 Mb, where one set of chromosomes is inherited maternally while the other set is inherited paternally[11]. Chromosomes 1 - 22 form homologous pairs and are known as *autosomes*, while chromosome pair 23 is partially

homologous and referred to as the *sex chromosome*.

There are two types of sex chromosomes, the X chromosome and the Y chromosome. These chromosomes differ in size and the genes they contain. The X chromosome spans more than 155 Mb while the Y chromosome only spans 59 Mb; representing a little under 5% and 2% of the total genome length, respectively[12]. Females have two homologous X chromosomes while males have one X chromosome and one Y chromosome. Due to the different numbers of each sex chromosome between males and females, their inheritance pattern differs to that of the autosomes.

During sexual reproduction, the autosomes undergo meiosis as described in section 1.1.1, producing offspring that inherit one set of recombined autosomal chromosomes from their mother and another set from their father (Figure 1.3). Similarly, the X chromosomes of females undergo meiosis such that all offspring inherit a single maternally-recombined X chromosome. The sex chromosomes of males however, undergo meiosis in a slightly different manner. Here, recombination only occurs between the homologous regions of the X and Y chromosomes, the *pseudoautosomal* regions, which encompass a small proportion of the sex chromosomes (<2% of the X chromosome and <5% of the Y chromosome[12]). As such, the sex chromosomes of males remain largely unaltered during meiosis. Therefore daughters inherit their second X chromosome as an almost identical copy from their father, while sons inherit their fathers Y chromosome as a close replica (Figure 1.3).

We focus on the autosomes and the X chromosome throughout this thesis. This includes the X chromosomes of both males and females, even though a male's X chromosome does not recombine, as recombination can occur in future generations. Furthermore, we exclude pseudoautosomal regions from our work on the X chromosome.

1.2 Identity by descent

1.2.1 Background

Two alleles are *identical by state* (IBS) if they have the same nucleotide sequence. These alleles can be further classified as *identical by descent* (IBD) if they have been inherited from a common ancestor[13]. Thus, alleles that are IBD must also be IBS, however the converse of this statement is not true.

Given that alleles must be inherited from a common ancestor in order to be IBD, it follows that IBD regions are only observed in related individuals. Individuals who are

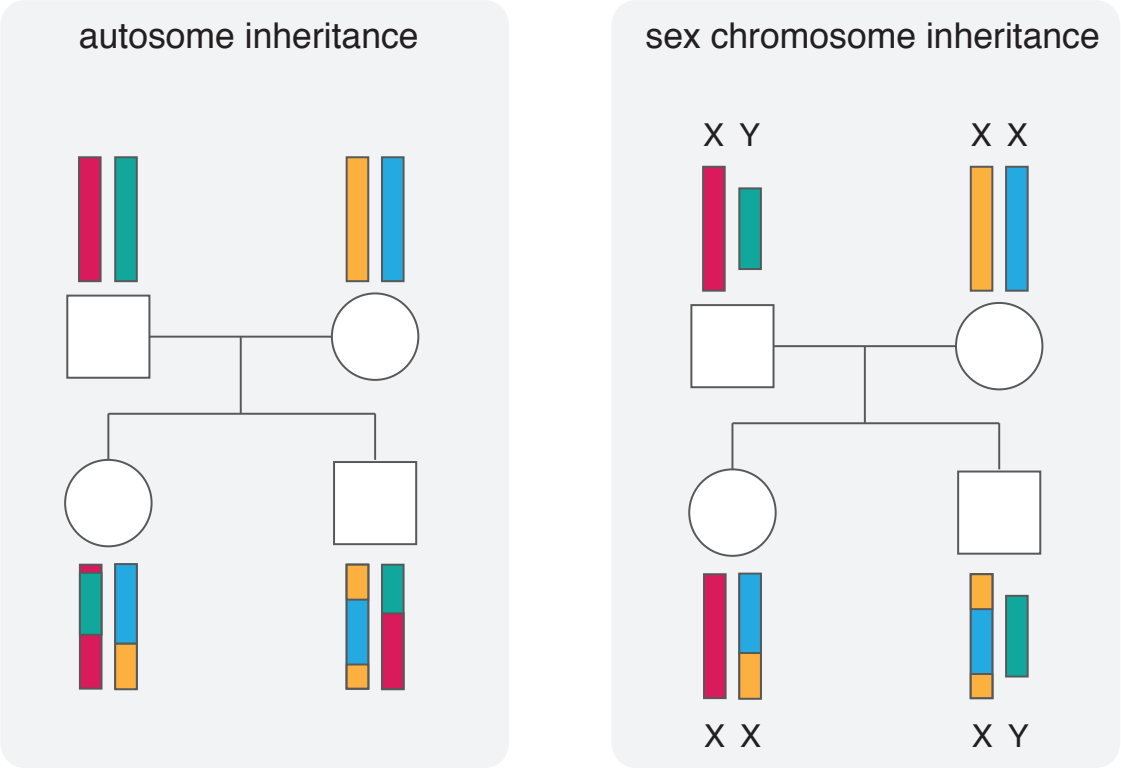


Figure 1.3: Two generation pedigrees displaying the inheritance pattern of autosomes and the sex chromosomes in the human genome. Females are denoted by circles and males by squares, with founders at the top of the pedigrees and offspring below them. The haplotypes are represented by coloured blocks above the founders and below the offspring. This figure was adapted from Griffiths et al.[1] and drawn using Adobe Illustrator.

closely related tend to share a large proportion of their genome IBD with many segments of considerable size. As the relationships become more distant, the amount of genome shared IBD decreases, as do the lengths of the respective IBD segments. This is the result of recombination breaking the segments into smaller fragments over multiple generations.

The expected proportion of genome shared IBD between two individuals and the respective lengths of the IBD segments can be estimated if the number of meiosis, m, separating the pair of individuals is known[14,15]. The average proportion of genome shared IBD, i.e. the sum of all IBD segments divided by the genome length, is calculated as $2^{-(m-1)}$. Here, the proportion of genome IBD is halved for each additional meiosis, as the probability of an allele being inherited on a single homologous chromosome is 50%. In contrast, the average length of each IBD segment decreases exponentially as the number of meiosis increases, with a mean value 100^{-m} cM. For example, third cousins are separated by eight meioses and share on average 0.8% of their genome IBD (24 cM or $\sim$ 24 Mb) with IBD segments spanning 12.5 cM ($\sim$ 12.5 Mb) on average.

Chromosomal inheritance patterns are slightly more complicated for the X chromosome

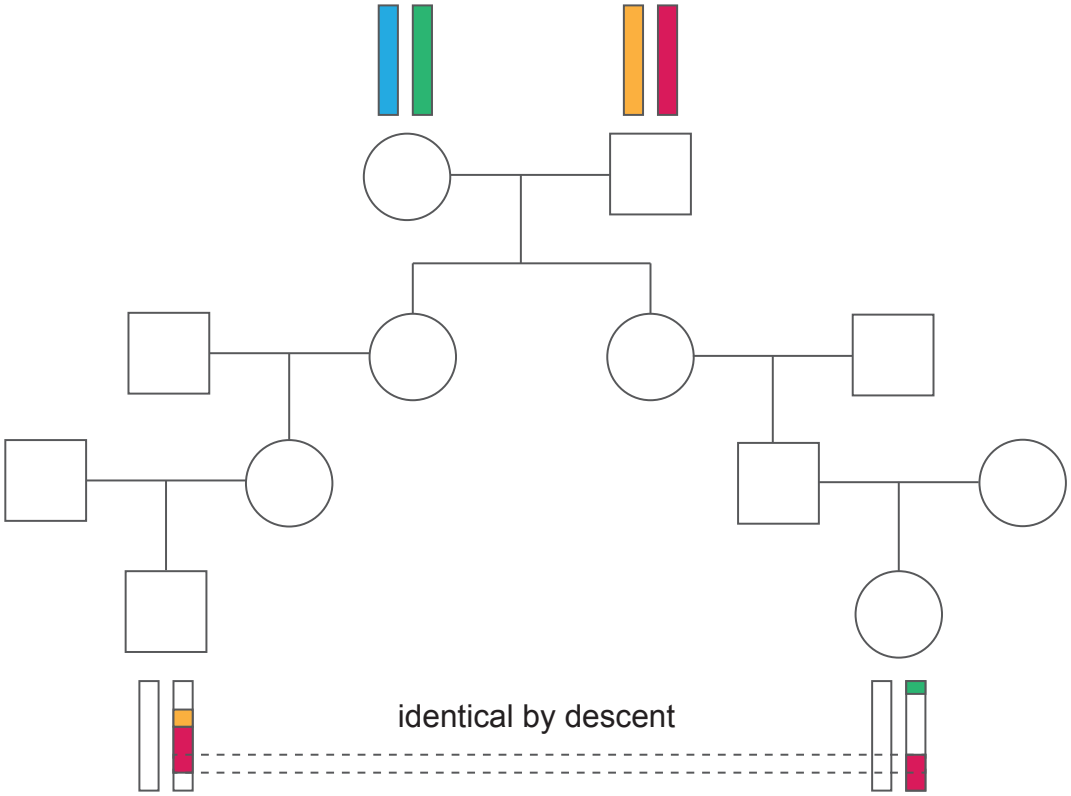


Figure 1.4: A four generation pedigree displaying a genomic region that is IBD between second cousins. The IBD segment has been inherited from their great-great grandfather.

than for autosomes and as such, the genders of the individuals in the lineage separating a pair of individuals, and the order in which these individuals appear, becomes an important factor in determining where or not two individuals can be related through the X chromosome. This is simply because only daughters can inherit their fathers X chromosome. Therefore, it is impossible for two individuals to share segments of IBD on the X chromosome in lineages with two or more male transmissions in earlier generations (assuming there is no consanguinity). At most, every second individual in an X chromosome lineage can be male for IBD to be possible, while the number of female-female transmission is unrestricted. This gives rise to the potential for segments of IBD on the X chromosome to be larger than those inherited on autosomes, simply because the male X chromosome does not undergo recombination (excluding pseudoautosomal regions), resulting in one less meiosis event for every male transmission in an X chromosome lineage, excluding initial founders.

As individuals become more distantly related, the chance of inheriting an IBD segment diminishes substantially, for both the autosomes and the X chromosome[14,15]. However, very small segments of IBD tend to persist are inherited from extremely distant com-

mon ancestors. Such IBD is the result of non-random allele associations over multiple loci[16], which is commonly referred to as *linkage disequilibrium* (LD). LD that has resulted from ancient ancestry is reflective of population substructure, whereby individuals within populations are more closely related than those in different populations[16,17]. Although interesting in its own right, identifying ancient relatedness due to population LD is not focus of this thesis. Instead, we are concerned with identifying recent common ancestry. Therefore we impose a timeframe on IBD such that only recent ancestry up to 25 generations (50 meioses) will be examined. In doing so, the smallest segment that we hope to identify in individuals separated by up to 25 generations will span 2cM on average[15].

Identifying genomic regions shared IBD from recent common ancestry is the basis of relatedness mapping. Such regions have proven useful in many applications, including disease mapping[15,18,19], identifying unknown relatedness[20] and detecting natural selection[21,22], among other things. We expand on these applications below.

1.2.2 Disease mapping

IBD analyses have been extensively performed in disease mapping, which aims to localize a critical region containing disease susceptibility genes[18]. This is done by first establishing some form of relatedness between affected individuals, in which case the disease gene is likely to have been inherited, and hence IBD. Thus critical regions are simply the IBD regions that segregate with the affected individuals.

When many affected individuals are analyzed with the same disease susceptibility gene an abundance of IBD is often observed, producing a genomic signal that readily identifies the critical region. Albrechtsen et al.[18] performed an IBD analysis on seven purportedly unrelated individuals who have an identical causal variant for breast and/or ovarian cancer in the *BRCA1* gene. Distant relatedness was discovered between the seven individuals and excessive IBD was observed in a region containing the *BRCA1* gene. Similarly, Browning and Browning[15] performed an IBD analysis of two supposedly unrelated families with multiple cases of gray platelet syndrome (GPS), an extremely rare inherited bleeding disorder, whereby both families shared an identical variant causal for the disease in the gene *NBEAL2*. They too discovered relatedness between the families and were able to identify a critical region of 7.7 Mb containing the disease gene.

Prior to the discovery of the disease gene *NBEAL2*, a linkage analysis of six unrelated families with GPS, excluding the two families analysed in Browning and Browning[15], had

identified a 9.3 Mb critical region containing *NBEAL2*[23]. This analysis was performed using 14 affected individuals while Browning and Browning were able to better refine the critical region using IBD analysis of only four affected individuals. Furthermore, IBD analyses of as little as two affected individuals have also successfully identified disease critical regions. Shaw et al.[19] were able to confirm that a novel variant in the gene *L1CAM*, found in two purportedly unrelated individuals with X-linked intellectual disability, was indeed causal for the disorder through identification of a 5.6 cM IBD segment containing the gene.

Albrechtsen et al.[18], Browning and Browning[15] and Shaw et al.[19] successfully localized the disease critical region using IBD analysis, which was enhanced by the discovery of distant relatedness. Analyses containing such distant relatives can better refine disease critical regions as these individuals share less of their genome IBD with other individuals and the segments that are shared have shortened over multiple generations. Linkage analysis also uses IBD segments to determine critical regions[7], however it requires knowledge of an accurate pedigree, which is not always possible for distant relatives. While linkage analysis is a powerful tool for discovering disease critical regions in Mendelian disorders[24,25], analysis of large pedigrees or distant relatives bears computational challenges and quickly becomes intractable[26,27].

Pedigree-based IBD analyses, such as linkage analysis, are concerned with very recent common ancestry while non-pedigree analyses, including IBD mapping, are useful for more distant relatedness[15,28]. This can be extended to extremely distant ancestry, or ancient ancestry, which is the primary focus of genome wide association studies (GWAS). GWAS are concerned with finding extremely short segments of IBD that reflect population LD[29]. Unlike IBD mapping or linkage analysis, which favors identification of genes with rare variants at moderate to high penetrance[25,30,31], GWAS are designed to identify genes with common variants at low penetrance[32]. In order to identify such variants, GWAS require a cohort of hundreds to thousands of individuals with the same trait and a matched control cohort for comparison. While GWAS can be extremely powerful for identifying disease genes[29], such sizable datasets are not always available, particularly for rare diseases, which make them impractical in many instances.

1.2.3 Identifying recurrent variants

Unlike linkage analysis and GWAS, the individuals contributing to an IBD signature at a specific locus can be extracted using IBD mapping. This is advantageous as it allows for the distinction between a variant that is likely to be *recurrent*, i.e. a variant that has arisen independently within a species on multiple haplotype backgrounds, and that which has come from a common ancestor[30,33].

1.2.4 Quantifying relatedness

The amount of genome shared IBD and the lengths of the respective segments vary depending on the number of generations separating two individuals. As such, IBD can be used to verify, or potentially refute, purported relationships as well as identify relatedness between supposedly unrelated individuals. Pemberton et al.[20] performed an IBD analysis of the HapMap Phase III data and were able to confirm most of the previously reported relationships. They identified four instances where relationships had been misclassified as well as an additional 177 closely related individuals who were thought to be unrelated, among which were parent-offspring pairs, full siblings and half siblings, avuncular and even monozygotic twins.

Recently, IBD analyses have become available to the general public through commercial genomics companies such as 23andMe (www.23andMe.com) and AncestryDNA (www.ancestory.com), which allow consumers to explore their heritage via ancestry reports.

1.2.5 Identifying Selection

Natural selection occurs when genetic traits that improve the survival of a species are passed on to offspring such that the traits increase in prevalence in the population[1]. In particular, traits that are beneficial to a species, and thus selected for, undergo *positive selection*. For very recent positive selection, an allele that is selected for has typically undergone few recombination events and is located on longer than expected haplotypes that are easily identified by IBD analysis[21,34]. As such, IBD analysis has proven to be a powerful tool for detecting very recent and very strong positive selection[21,22] and we detail several instances of recent positive selection here (Figure 2).

Positive selection can occur as a *hard sweep*, whereby an allele introduced into a population reaches a high frequency over a short period of time[34]. A hard sweep is generally

characterized by an increase in frequency of shared segments of considerable size over the favored allele, which produces a strong genetic signature that is readily identified by IBD algorithms. When the allele reaches fixation in a population, variability from mutations and recombination are introduced and the signal dissipates.

Albrechtsen et al.[21] performed an IBD analysis of the HapMap phase III dataset and discovered the human leukocyte antigen (HLA), involved in the immune systems response to foreign invaders, to be positively selected in eleven of twelve populations studied. Han and Abney[22] performed an IBD analysis on individuals from Kenya and also identified HLA to be under positive selection in addition to the human lactase gene (LCT), which is responsible for making the enzyme lactase that enables digestion of milk post infancy and is most commonly selected for in individuals of European descent.

A more complicated type of recent positive selection is *selection on standing variation*, whereby an allele that is already present in the population at a modest frequency suddenly becomes favored[34]. Since the favored allele was already present in the population, neighboring variants have time to recombine and associate with different background haplotypes. This results in shorter segments of IBD than a hard sweep, which are more difficult to identify with IBD analysis, although the frequency of shared segments typically remains high. Albrechtsen et al.[21] performed simulation studies to compare the genomic signatures of a hard selective sweep and selection on standing variation. They found the signatures to be more pronounced for selection on standing variation soon after selection of the favored allele, however these signatures dissipated more rapidly than with a hard selective sweep.

The last type of recent positive selection that we consider leaves a more subtle, and therefore harder to identify, signature in the genome[36]. This type of positive selection is a *soft selective sweep*. Here recurrent variants on different haplotype backgrounds increase in frequency in the population[34]. A greater number of haplotype backgrounds leads to fewer pairs with IBD sharing and diluted signals of excess IBD[21], compared to a hard sweep or selection on standing variation. When the selected alleles are at low frequencies in the population the signature of positive selection may be non-existent. Thus a considerable number of generations may need to have passed before a signature is detected. The ability to identify a soft selective sweep via IBD depends on the number of haplotype backgrounds and the frequency of these haplotypes in the population.

Figure 1.5: Three ways in which an allele can be positively selected, adapted from Scheinfeldt et al.[35]. Each line represents a haplotype and colored circles are unique SNPs on the haplotype. Black and white circles denote the alleles under selection. **A.** A hard sweep occurs when an allele is introduced into a population and increases in frequency rapidly such that the surrounding haplotype also increases in frequency due to LD and few recombination events. **B.** Selection on standing variation occurs when an allele that is already present in the population (grey circle) at moderate frequency becomes selected for and rapidly increases in frequency, along with neighboring SNPs. **C.** A soft selective sweep occurs when the same variant (back and white circles) arises on multiple haplotype backgrounds, and each haplotype increases in frequency, where a single haplotype need not dominate.

1.3 SNP technologies

SNPs have gained popularity in recent years as markers of genetic diversity and have been extensively used to identify the genetic causes and predisposition of disease[1]. They are also the foundation of IBD analysis. Large segments of IBD tend to contain many SNPs with identical alleles and are easily identified[27]. As SNP technologies continually allow for increasing SNP density, smaller IBD segments are also becoming more easily identified. However this comes at the cost of incorrectly identifying IBD from extremely ancient ancestry, that is the result of population LD, as IBD from more recent common ancestry[17,27]. Furthermore, increased SNP density also results in increased computational

demand, not only from the sheer increase in size of the datasets analyzed but also from models now having to account for LD. Although somewhat challenging, these issues can be overcome by clever computing algorithms and segment filtering.

The identification of IBD segments using SNPs has been made possible by advances in technologies leading to the development of genotyping arrays and next generation sequencing (NGS). Genotyping arrays target hundreds of thousands to millions of common SNPs in the genome, usually within genes. Simply, genotyping arrays work by hybridizing fragments of sample DNA to fluorescently tagged DNA probes. Genotypes can then be extracted for SNPs from fluorescence intensity profiles[37]. Such arrays allow for many different variants to be examined at once and are extremely cost effective. They do, however, require prior knowledge of the SNPs of interest and may not be suitable for identifying the causal variant of extremely rare diseases (assuming the causal variant is a SNP).

NGS has become extremely popular in recent years, as it enables the whole genome, or targeted regions of the genome, to be sequenced at continually decreasing costs. NGS works by fragmenting the sample DNA into smaller segments then aligning these segments to a reference genome or performing de novo assembly[38]. SNPs are then called at loci that differ to the reference genome. Genome sequencing allows for the exact sequence of a DNA segment to be identified, which can be used to find known SNPs as well as SNPs that are unique to the sample.

Both genotyping arrays and NGS suffer from occasional genotyping errors, with higher rates observed in NGS data. This can affect the identification of IBD when the resultant SNP switches between homozygous reference and homozygous alternative (or visa versa), or homozygous to heterozygous (or visa versa), resulting in allele sequences that are inconsistent with IBD and are less likely to be identified as such.[39] performed analyses using simulated SNP array data and showed that genotyping error rates varying between 0% and 0.5% had little impact of IBD inference in disease mapping. This is reassuring for IBD inference. Alternatively, rapid switching between IBD and non-IBD in small genomic intervals can be used to detect genotyping errors[40].

1.4 IBD methodologies

Existing tools for performing IBD mapping differ in many ways. Browning and Browning[27] give a nice overview of the methodologies and we expand on this here.

Algorithms can more broadly be classified as probabilistic or non-probabilistic models. Probabilistic algorithms include hidden Markov models[17,18,27,41,42,43,44,45,46] (HMM; described in detail in Chapter 2) while non-probabilistic algorithms commonly use thresholds on IBS consistent haplotypes to infer IBD[47,48,49]. Probabilistic models are generally preferred to non-probabilistic models as they have the ability to account for LD, which has become crucial as the density of SNPs continues to increase with advances in genotyping technologies. Algorithms can also differ in the way they model IBD status. Binary states[27,43] (IBD vs non-IBD) and ternary states[18,41] (the number of alleles shared IBD from either 0, 1 or 2 alleles) are commonly used to model IBD status. Less commonly used are are Jacquard's 15 identity coefficients for phased data[17,46] and Jacquard's 9 condensed identity coefficients for unphased data[17,44,46,50]. In addition to this, some tools are designed primarily for family based studies where accurate pedigree information is required[42,44] while other tools specialize in population-based cohort studies where relatedness between individuals is often unknown. We describe in detail some of the well used tools available for inferring IBD that have adopted the more popular approaches and provide a more extensive list in Tables 1.1, 1.2 and 1.3.

1.4.1 Thresholding approaches

Some of the earlier IBD methodologies were simple non-probabilistic methods that inferred IBD based on IBS information[33,57,58,59]. These methods used genotype data to find stretches of the genome that were IBS between pairs of individuals and called the region IBD if its length was larger than a length threshold determined by the number of generations separating the individuals. Since the expected length of an IBD segment decreases as a pair of individuals become more distantly related, accurate pedigree information is typically required to determine the length threshold used in these algorithms. Unfortunately however, pedigree information is not always available.

This problem was eliminated by GERMLINE[47], which also uses a length threshold constraint on IBS consistent haplotypes to determine IBD regions. GERMLINE implements a sliding window approach where a window of predefined length moves along the genome partitioning it into non-overlapping bins. A dictionary of allele combinations for all samples is created for each bin, from which IBS is determined, allowing for occasional genotyping errors. GERMLINE does not require specification of a pedigree so can be used on individuals with unknown relatedness. However, it does require the window size (in

Table 1.1: Tools for inferring IBD segments

Tool	Author	Algorithm description	Haploid[a]	Unphased data[b]	No pedigree required	LD model[c]	Genotyping errors[d]	Missing data[e]
PLINK	Purcell et al.[41]	Three state HMM for unphased genotype data in LE	✗	✓	✓	✗	✗	✗
GERMLINE	Gusev et al.[47]	Sliding window with length threshold on IBD consistent haplotypes	✓	✓	✓	✗	✓	✗
RELATE	Albrechtsen et al.[18]	Three state HMM for unphased genotype data with conditional emission probabilities to account for LD	✗	✓	✓	✓	✓	✓
IBDfinder	Carr et al.[48]	Performs IBS SNP-counting to identify homozygosity by descent regions with large scores.	✗	✓	✓	✗	✓	✗
BEAGLE IBD	Browning and Browning[27]	Two state HMM with localized haplotype cluster model for LD	✗	✗	✓	✓	✓	✓
IBDmap	Bercovici et al.[42]	A factorial HMM that uses a first order model to account for LD in pedigrees	✗	✓	✗	✓	✗	✓
ibd2	Krawitz et al.[43]	Two state HMM to infer homozygosity by descent	✗	✓	✗	✗	✓	✗

[a] IBD can be inferred on haploid chromosomes. **NOTE:** All algorithms can analyze haploid chromosomes if the chromosomes are duplicated. However probability distributions may be incorrect and as such we do not consider algorithms with probability distributions tailored for diploid chromosomes to be applicable to haploid chromosomes.
[b] IBD is detected from unphased data such that phased data is **not** required by the algorithm and phasing is **not** performed as part of the analysis.
[c] A model for LD is included in the algorithm.
[d] A model for genotyping errors is included in the algorithm.
[e] SNPs with missing data can be included in the analysis.

Table 1.2: Tools for inferring IBD segments continued...

Tool	Author	Algorithm description	Haploid[a]	Unphased data[b]	No pedigree required	LD model[c]	Genotyping errors[d]	Missing data[e]
fastIBD	Browning and Browning[49]	Sliding window with haplotype frequency threshold on IBD consistent haplotypes and localized haplotype cluster model for LD	X	X	✓	✓	✓	✓
IBDLD	Han and Abney[44]	Nine state HMM with ridge regression to account for LD between multiple loci	X	✓	X	✓	✓	✓
MCMC_IBDfinder	Moltke et al.[51]	Infers IBD in multiple individuals simultaneously using Markov Chain Monte Carlo	X	✓	✓	X	✓	X
IBD_Haplo	Brown et al.[17]	A nine (or 15) state HMM for phased (unphased) data	X	✓	✓	X	✓	✓
Refined IBD	Browning and Browning[52]	Uses GERMLINE to infer IBD then uses the BEAGLE HMM to calculate a likelihood ratio (LOD score) for an IBD vs non-IBD model	*	X	✓	✓	X	✓
IBDseq	Browning and Browning[53]	Likelihood ratio of observed genotypes for IBD vs non-IBD model in sequencing data	X	✓	✓	X	✓	✓

[a] IBD can be inferred on haploid chromosomes. **NOTE:** All algorithms can analyze haploid chromosomes if the chromosomes are duplicated. However probability distributions may be incorrect and as such we do not consider algorithms with probability distributions tailored for diploid chromosomes to be applicable to haploid chromosomes.
[b] IBD is detected from unphased data such that phased data is **not** required by the algorithm and phasing is **not** performed as part of the analysis.
[c] A model for LD is included in the algorithm.
[d] A model for genotyping errors is included in the algorithm.
[e] SNPs with missing data can be included in the analysis.

Table 1.3: Tools for inferring IBD segments continued...

Tool	Author	Algorithm description	Haploid[a]	Unphased data[b]	No pedigree required	LD model[c]	Genotyping errors[d]	Missing data[e]
IBD-Groupon	He[45]	Uses fastIBD to infer pairwise IBD then uses a HMM to select most likely groupwise IBD	✗	✗	✓	✓	✓	✓
HapFABIA	Hochreiter[54]	Performs biclustering to infer very short IBD segments in multiple individuals that are tagged by rare variants	✗	✓	✓	✗	✓	✗
Parente2	Rodriguez et al.[55]	Calculates the log-likelihood ratio of IBD in overlapping sliding windows	✗	✗	✓	✓	✓	✗
ibd_stitch	Glazner and Thompson[46]	Uses a HMM to infer pairwise IBD then builds a joint IBD graph to infer IBD in multiple individuals	✗	✓	✓	✗	✓	✓
ExIBD	Fu et al.[56]	Uses fastIBD to infer pairwise IBD in exome sequencing data then refines segments with BEAGLE IBD	✗	✗	✓	✓	✓	✓

[a] IBD can be inferred on haploid chromosomes. **NOTE:** All algorithms can analyze haploid chromosomes if the chromosomes are duplicated. However probability distributions may be incorrect and as such we do not consider algorithms with probability distributions tailored for diploid chromosomes to be applicable to haploid chromosomes.
[b] IBD is detected from unphased data such that phased data is **not** required by the algorithm and phasing is **not** performed as part of the analysis.
[c] A model for LD is included in the algorithm.
[d] A model for genotyping errors is included in the algorithm.
[e] SNPs with missing data can be included in the analysis.

SNPs) to be specified which can influence the ability to infer IBD segments. Too large a window size can result in real IBD segments being missed while too small a value results in false positive calls. While selecting the window size can cause problems, the main advantage of GERMLINE is its computational power. GERMLINE can perform analyses on thousands of individuals using dense genotype or phased haplotype data in a matter of seconds, making it a very popular method for population-based cohort studies.

Browning and Browning adopted GERMLINE's computationally efficient haplotype-dictionary approach in two of their IBD algorithms. Initially, Browning and Browning[49] developed fastIBD, which determines IBD from haplotype frequency rather than segment length. fastIBD uses unphased genotype data to create a localized haplotype cluster model[60] from which multiple phased haplotypes are sampled for each individual. These haplotypes are themselves used to create a localized haplotype cluster model that defines a HMM; the BEAGLE HMM[61]. A sliding window is used to search for identical haplotypes traversing the same path in the BEAGLE HMM that are then identified as IBD if their haplotype frequency is less than a predefined threshold. Unlike GERMLINE, fastIBD models LD through the BEAGLE HMM. However a sufficient number of individuals are required to build the HMM which is not always available. Additionally, fastIBD does not allow for genotyping errors, which may be problematic as the density of markers in SNP arrays increases resulting in more genotyping errors. Furthermore, the accuracy of fastIBD depends on the ability of the phasing algorithm that is implemented within the tool, where poorly phased data containing genotyping errors could result in misleading IBD calls.

Following this, Browning and Browning[52] developed Refined IBD which initially uses GERMLINE to find candidate IBD segments then applies a probabilistic approach to assess the evidence of IBD and refine the candidate list. In the refinement step, phased haplotypes are used to build the BEAGLE HMM and the likelihood of one haplotype shared IBD and no haplotypes shared IBD are calculated for each candidate segment. LOD scores are then calculated using the likelihood ratios and tracts with LOD scores less than a user defined threshold are removed. The additional step allows for higher accuracy than GERMLINE, however as with fastIBD the accuracy of this method depends on the ability of its built-in phasing algorithm, and genotyping errors are not accounted for.

More recently, Rodriguez et al.[55] developed Parente2, which slides a block of user-defined length (in cM) beginning at each marker across each chromosome. The blocks

are subset into smaller windows of possibly non-consecutive SNPs, from which scores are calculated based on the log likelihood ratio (LRR) of an IBD vs non-IBD model. Window scores are summed for each block and a block is called IBD if its score is greater than a predefined threshold. Rodriguez et al.[55] have demonstrated that Parente2 is more computationally efficient than GERMLINE and fastIBD in addition to it accounting for LD. Parente2 accounts for LD by explicitly modeling haplotype frequencies in the LLR from a phased training dataset that appropriately match the input dataset. Unfortunately, such datasets are not always available.

1.4.2 Hidden Markov model

One of the first probabilistic IBD methodologies was PLINK[41]. PLINK implements a continuous time HMM on two diploid chromosomes to infer whether 0, 1 or 2 alleles are shared IBD between pairs of individuals using genotype data. PLINK can be applied to cohorts of unrelated individuals and does not require any user specified parameters. Unfortunately however, PLINK requires the genotype data to be in approximate LE which often involves thinning SNPs prior to use. Thinning SNPs can result in a substantially smaller dataset to analyze, and in some instance less than 5% of the original data may remain. Reducing datasets like this can result in the loss of potentially informative SNPs, however in doing so PLINK becomes computationally efficient and comparable to some of the fastest algorithms designed for cohorts with thousands of individuals. PLINK's algorithm requires population allele frequencies which are calculated from the input dataset. Frequency bias can arise from this if the input dataset is too small or the individuals are of mixed ethnicities.

Albrechtsen et al.[18] extend the HMM proposed by PLINK to allow for LD, genotyping errors and missing data. This method was implemented in RELATE, which accounts for LD through the emission probabilities where the current genotype probability is conditioned on the genotype of a single previous marker. Since RELATE allows for LD there is no need to prune the dataset like PLINK, however all LD cannot be accounted for in very dense datasets using this algorithm which can result in false positive IBD calls. Allele frequencies and haplotype frequencies can be calculated from a reference dataset if available, otherwise these values are calculated from the input dataset which may not reflect true population frequencies for small datasets.

Prior to fastIBD and Refined IBD, Browning[62] implemented a HMM to model binary

IBD status between pairs of haplotypes, which was then extended in BEAGLE IBD to identify IBD between pairs of diploid individual[27]. BEAGLE IBD first builds a localized haplotype cluster model to phase genotype data and model LD using the BEAGLE HMM; it then adds an IBD model to calculate the posterior probability of IBD for each marker. Regions with probabilities above a threshold are considered IBD. BEAGLE IBD is computationally intensive and, as BEAGLE continually releases improved IBD methodologies, has been superseded by BEAGLE fastIBD and BEAGLE Refined IBD.

Han and Abney[44] further extended the models of Purcell et la.[41] and Albrechtsen et al.[18] to better account for LD in IBDLD. IBDLD implements a nine-state HMM for Jacquard's condensed identity coefficients[50] and accounts for LD by conditioning on multiple marker genotypes using a linear model with ridge regression. Although IBDLD implements a more polished model for LD than Albrechtsen et al.[18], IBDLD requires relationships between the individuals must be known.

1.4.3 Tools for NGS data

With the advent of NGS, algorithms for IBD detection are now exploring how to better account for increasing SNP density and higher genotyping error rates, as well as the genomic patchiness that can be produced with targeted sequencing. Browning and Browning[53] implemented a likelihood ratio approach in IBDseq for unphased sequencing data. Unfortunately, IBDseq requires LD thinning as LD is not directly accounted for in the model. Despite this, IBDseq achieves high power and accuracy in sequencing data and is computationally efficient for analyses with many SNPs.

Fu et al.[56] have recently developed ExIBD, an algorithm for exome sequencing data which infers IBD segments using fastIBD then implements BEAGLE IBD to refine candidate segments. These algorithms were developed for array data, however when used together in ExIBD, perform relatively well on exome-sequencing data, excluding regions with low exon density.

While these tools were designed specifically for use with sequencing data, we note that all algorithms can be applied to sequencing data once SNPs have been extracted, although filtering procedures may be required to reduce false positive IBD detection that is the result of population LD. Additionally, some algorithms may be computationally inefficient with many SNPs and are better suited for array data. Furthermore, non-uniformly spaced SNPs may result in undetected IBD segments.

1.4.4 IBD methodologies incorporating haploid chromosomes

Few software can perform IBD analyses on haploid chromosomes. Of all the methodologies in Tables 1.1, 1.2 and 1.3, only GERMLINE can be implemented on haploid chromosomes as it does not involve any probability distributions. All other methodologies have been designed for diploid chromosomes, and while haploid chromosomes can be duplicated to give the appearance of a diploid chromosome, the probability distributions governing the algorithms do not correctly account for this. For example, fastIBD uses the localized haplotype cluster model in the BEAGLE HMM to generate phased data and calculate haplotype frequencies from diploid chromosomes. While a haploid chromosome is trivially phased and will not be altered by the BEAGLE phasing algorithm, the haplotype frequencies in the model will be incorrect if such chromosomes are duplicated to form a pair of homozygous chromosomes. In a dataset containing 20 haploid chromosomes and 20 diploid chromosomes, the haploid chromosomes will contribute 50% (40/80) to the haplotype frequency calculations if duplicated, whereas they should only contribute 33% (20/60).

Similarly, tools such as PLINK and RELATE, which implement three-state HMMs to determine whether 0, 1 or 2 alleles are shared IBD at each SNP, have incorrect probability distributions if haploid chromosomes are duplicated to form a pair of homozygous chromosomes. Here the emission probabilities are functions of the population allele frequencies, the observed genotype for a pair of diploid chromosomes and the state space. If a pair of diploid chromosomes each have the genotype AA at the same SNP, where A is the reference allele, then the probability of observing the pair of genotypes $\{AA, AA\}$ given there is no IBD sharing is p_A^4, where p_A is the population reference allele frequency. If one of these chromosomes is haploid, then the probability of observing the pair of genotypes $\{AA, A\}$ given there is no IBD sharing is p_A^3, however PLINK and RELATE are unaware that one chromosome is haploid and still calculate the probability as p_A^4.

While one methodology does exist to correctly perform IBD analyses between haploid chromosomes, the algorithm is nonprobabilistic and as such it does not account for LD, which has become an important criterion in an IBD model with increasing SNP density from NGS data. As such, a probabilistic model is required that can correctly analyse haploid chromosomes as well as account for LD.

1.5 Aims of study

This thesis describes a novel algorithm for inferring pairwise IBD using SNP genotype data that correctly accounts for haploid and diploid chromosomes. Chapter 1 serves as a general introduction into common sources of genetic variation, including recombination during meiosis and SNPs, and describes how recombination influences genetic relatedness. This chapter describes applications that take advantage of relatedness through IBD, and reviews the current methodologies for IBD detection.

Most algorithms for inferring IBD are for analysis of diploid genomes. This typically results in exclusion of the X chromosome from analysis as it is diploid in females and haploid in males. Therefore, in Chapter 2 we detail an algorithm for inferring IBD on both diploid and haploid chromosomes. It is necessary to extensively examine the performance of new algorithms. As such, we perform statistical and bioinformatics analyses on simulated data to demonstrate the power and accuracy of this model. This work is presented in Chapter 3 and aims to fill a void in what is lacking from IBD methodologies.

Chapter 4 presents an application of the IBD methodology described in Chapter 2 to a small cohort of individuals affected with familial adult myoclonus epilepsy (FAME); a rare form of epilepsy with causal variants yet to be determined. This analysis investigates two critical regions that were identified using linkage analysis, and provides novel insights for future investigations of this disorder.

Following applications and analysis of the human genome, we modify our algorithm such that it can be applied to non-human organisms with haploid genomes. We are specifically interested in microorganisms that cause infectious disease, such as the malaria causing parasite, *Plasmodium*, with the aim of identifying loci under positive selection using IBD. The shift in focus to haploid microorganisms comes amid the global antimicrobial resistance crisis, whereby a number of microorganisms have become resistant to first-line and/or last-resort antimicrobial treatments, threatening to increase disease incidence and mortality rates (WHO, 2015). Therefore, identifying the genomic mechanisms underlying antimicrobial resistance is essential to reduce the burden of this crisis.

The remainder of this thesis delves into the disease, malaria, and the parasite responsible for this disease, *Plasmodium*. Chapter 5 presents an introduction into the biology of the parasite and a brief history on antimalarial drug resistance. We discuss the popular tools that are used to identify loci under selection associated with resistance and what is

required from an IBD methodology in order to identify such loci in *Plasmodium*.

Chapter 6 describes the modifications of the algorithm in Chapter 2 that are required for IBD analysis of *Plasmodium*, then performs an in-depth analysis of a global malaria dataset. This analysis uncovers novel insights into the disease, including antimalarial drug resistance, and demonstrates the importance of IBD analysis of disease-causing microorganisms.

Finally, Chapter 7 serves as a general discussion of the IBD methodology presented in this thesis and the importance of its development for both human and non-human studies. We discuss implications of our results, specifically with regards to the *Plasmodium*, and further work and applications in the wider field of IBD analysis and the infectious disease setting.

Chapter 2

XIBD: pairwise identity by descent methodology

This chapter details the XIBD methodology that was developed to detect IBD segments in both diploid and haploid chromosomes. The algorithms implemented are an extension of Purcell et al.[41] and Albrechtsen et al.[18], both of which developed hidden Markov models. As such, this chapter also provides an extensive overview of a hidden Markov model. This work has been published[63] and the XIBD tool is available as an open source R package (https://github.com/bahlolab/XIBD). The relevant manuscript has been included as Appendix A and we extend our description below.

2.1 An introduction to hidden Markov models

A Hidden Markov model (HMM) is a probabilistic model that describes a sequence of events which gave rise to a set of observations. These models have been used in speech recognition systems[64], predicting the financial market[65] and copy number variation analysis[66], among other things. In order to understand the mechanisms behind a HMM it is helpful to first understand a Markov model. We briefly discuss a Markov model below and its extension to a HMM, where the definitions and notations are summarized from Rabiner[64].

2.1.1 Markov model

A Markov model is a stochastic model that describes a randomly changing system[67]. At any time t, the system is assumed to be in a particular state, where the collection of all

possible N states determines the state space $S = \{S_1, S_2, \ldots, S_N\}$. To predict the most likely state at time t, information on the previous states of the system must be known such that

$$\Pr(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots, q_1 = S_l),$$

where $1 \leq j, i, h, l \leq N$. As t increases, the number of probabilities that need to be calculated increases rapidly so a simplifying assumption is made. We assume the system satisfies the *Markov property* that, given the present state, the future and past states are independent. i.e.,

$$\Pr(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots, q_1 = S_l) \approx \Pr(q_t = S_j | q_{t-1} = S_i) \tag{2.1}$$

This is a *first-order* Markov assumption since the state of the system at time t only depends on the state at time $t-1$. An n^{th}*-order* Markov assumption occurs when the state of the system at time t depends on the state at time $t-1, t-2, \ldots, t-n$. The right hand side of equation 2.1 is independent of time, meaning

$$\Pr(q_{t+1} = S_1 | q_t = S_2) = \Pr(q_t = S_1 | q_{t-1} = S_2).$$

This allows us to calculate the state *transition probabilities*, a_{ij}, which are simply the probabilities of transitioning between states in a single time step;

$$a_{ij} = \Pr(q_t = S_j | q_{t-1} = S_i), \quad 1 \leq i, j \leq N \tag{2.2}$$

where

$$a_{ij} \geq 0$$

$$\sum_{j}^{N} a_{ij} = 1.$$

The collection of state transition probabilities forms the models *transition matrix*, $A = \{a_{ij}\}$. In addition to a transition matrix, the system requires *initial state probabilities*, i.e., the probabilities associated with beginning the system in a particular state. We denote the set of initial probabilities as $\Pi = \{\pi_i\}$ where

$$\pi_i = \Pr(q_1 = S_i), \quad 1 \leq i \leq N. \tag{2.3}$$

Using both the initial state probabilities and the state transition probabilities we can calculate the joint probability of a sequence of states $Q = q_1 q_2 \cdots q_T$ from this system as

$$\begin{aligned}\Pr(Q|A,\Pi) &= \Pr(q_1)\Pr(q_2|q_1)\Pr(q_3|q_2)\cdots\Pr(q_T|q_{T-1}) \\ &= \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}.\end{aligned}$$

A system with states that obey the Markov assumption is called a *Markov model*, and the random sequence of states that result from the system is called a *Markov chain*.

2.1.2 Hidden Markov model

In contrast to a Markov model where the states are observed, a *Hidden Markov model* (HMM) has unobservable states and instead has observations that are probabilistic functions of the states. Following the notation above, let $S = \{S_1, S_2, \ldots, S_N\}$ donate the N states, and $Q = q_1 q_2 \cdots q_T$ denote the hidden sequence of states of length T that we wish to determine. Additionally, let M define the number of distinct observations produced from all states and let $V = \{v_1, v_2, \ldots, v_M\}$ represent the observation symbols in the model. In addition to the initial state probabilities Π and transition probability matrix A defined in equations 2.2 and 2.3, a HMM requires one more probabilistic measure to be complete, namely the *emission probabilities*. Emission probabilities are the probabilities associated with the observations from the states, $B = \{b_j(k)\}$, where

$$b_j(k) = \Pr(v_k \text{ at } t | q_t = S_j), \quad 1 \leq j \leq N, \quad 1 \leq k \leq M. \tag{2.4}$$

As with Rabiner we let

$$\lambda = (A, B, \Pi)$$

denote the set of model parameters. In summary, a HMM is characterized by the following components:

1. The state space $S = \{S_1, \ldots, S_N\}$ comprised of N states.
2. M distinct possible observations that make up the observation set $V = \{v_1, \ldots, v_M\}$.
3. An observation sequence $O = o_1 \cdots o_T$ over T positions where $o_t \in V$.
4. The initial probability distribution $\Pi = \{\pi_i\}$ where

$$\pi_i = \Pr(q_1 = S_i), \quad 1 \leq i \leq N.$$

5. A state transition probability matrix $A = \{a_{ij}\}$ where

$$a_{ij} = \Pr(q_{t+1} = S_j | q_t = S_i) \quad 1 \leq i, j \leq N.$$

6. The emission probability distribution $B = \{b_i(k)\}$ where

$$b_j(k) = \Pr(v_k \text{ at } t | q_t = S_j), \quad 1 \leq j \leq N, \quad 1 \leq k \leq M.$$

2.1.3 Three basic problems of a HMM

According to Rabiner[64] there are three problems that need to be solved in order for an HMM to be useful in real-world applications;

1. How do we efficiently compute the probability of an observation sequence $O = o_1 o_2 \cdots o_T$, where $o_i \in V$, given the model $\lambda = (A, B, \Pi)$? i.e., calculate $\Pr(O|\lambda)$.
2. Given an observation sequence $O = o_1 o_2 \cdots o_T$ and the model $\lambda = (A, B, \Pi)$, how do we find a sequence of states $Q = q_1 q_2 \cdots q_T$, where $q_i \in S$, that best explains the observations?
3. How do we adjust the model parameters $\lambda = (A, B, \Pi)$ to maximize the probability of the observation sequence O? i.e., maximize $\Pr(O|\lambda)$.

Problem 1 The first problem of an HMM is how to efficiently calculate the probability of an observation sequence $O = o_1 o_2 \cdots o_T$, where $o_i \in V$, given the model $\lambda = (A, B, \Pi)$. That is, we wish to calculate $\Pr(O|\lambda)$. One way of calculating this probability is by

summing the joint probability of O and Q over all possible state sequences as follows:

$$\begin{aligned}
\Pr(O|\lambda) &= \sum_{\text{all } Q} \Pr(O, Q|\lambda) \\
&= \sum_{\text{all } Q} \Pr(O|Q, \lambda) \Pr(Q|\lambda) \\
&= \sum_{q_1, q_2, \ldots, q_T} \left(\Pr(q_1|\lambda) \Pr(o_1|q_1, \lambda) \prod_{t=2}^{T} (\Pr(q_t|q_{t-1}, \lambda) \Pr(o_t|q_t, \lambda)) \right) \\
&= \sum_{q_1, q_2, \ldots, q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) a_{q_2 q_3} \cdots b_{q_T}(o_T) a_{q_{T-1} q_T}. \qquad (2.5)
\end{aligned}$$

Computing $\Pr(O|\lambda)$ from equation 2.5 requires in the order of $2TN^T$ calculations and quickly becomes computationally demanding as T and N increase. Fortunately a computationally efficient method for calculating $\Pr(O|\lambda)$ exists, known as the *forward-backward algorithm*. The forward-backward algorithm avoids summing the probability of the observation sequence over an exponential number of paths Q, by calculating probabilities of reaching various states as time progresses as well as probabilities of finishing the sequence starting from a given state. These probabilities are called the forward probabilities and backward probabilities respectively and are derived below.

Let $\alpha_t(j)$ denote the forward probability given by

$$\alpha_t(j) = \Pr(o_1 o_2 \cdots o_t, q_t = S_j|\lambda).$$

We can compute $\alpha_t(j)$ for all t and j using induction as follows:

1. Initialization:

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N.$$

2. Induction:

$$\alpha_{t+1}(j) = \sum_{j}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}), \quad 1 \leq t \leq T-1, \quad 1 \leq j \leq N.$$

Since a_{ij} and $b_j(o_{t+1})$ are generally less than 1, we see that as t increases the forward

probabilities tend to zero. This is a problem as the computational precision required to calculate these probabilities will quickly exceed that available from most standard computers resulting in underflow. To avoid such problems, the forward probabilities are scaled using the following scaling coefficient c_t;

$$c_t = \frac{1}{\sum_{i=1}^{N} \alpha_t(i)},$$

where

$$\hat{\alpha}_t(i) = c_t \alpha_t(i) = \frac{\alpha_t(i)}{\sum_{i=1}^{N} \alpha_t(i)}.$$

The probability of the observation sequence $\Pr(O|\lambda)$ can be calculated from the scaling coefficients using

$$\Pr(O|\lambda) = \frac{1}{\prod_{t=1}^{T} c_t}.$$

However the product of the scaling coefficients is likely to result in underflow also so the log probability of the observation sequence is the preferred calculation;

$$\log(\Pr(O|\lambda)) = -\sum_{t=1}^{T} \log c_t.$$

Calculating the log probability of the observation sequence requires in the order of N^2T computations as apposed to $2TN^T$ from equation 2.5, which scales linearly rather than exponentially with N and T. The forward probabilities enable us to calculate $\Pr(O|\lambda)$ without having to calculate the backwards probabilities, however, the backwards probabilities are required to solve problem 2 so are defined below.

Let $\beta_t(i)$ denote the backwards probabilities given by

$$\beta_t(j) = \Pr(o_{t+1} \cdots o_T | q_t = S_j).$$

Similarly to the forwards probabilities, we can compute $\beta_t(j)$ for all t and j using induction:

1. Initialization:

$$\beta_T(i) = 1, \quad 1 \leq i \leq N.$$

2. Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, \ldots, 1, \quad 1 \leq i \leq N.$$

To avoid underflow, these probabilities are scaled using the same scaling coefficient as the forward probabilities. The scaled backwards probabilities are given by

$$\hat{\beta}_t(i) = c_t \beta_t(i) = \frac{\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)}.$$

Problem 2 The second problem of a HMM is, given an observation sequence $O = o_1 o_2 \cdots o_T$ and the model $\lambda = (A, B, \Pi)$, how do we find the most likely sequence of states $Q = q_1 q_2 \cdots q_T$, where $q_i \in S$, that could have generated O? As with problem 1, there is more than one solution to this problem and we describe two methods that are commonly used as a solution, namely the *posterior state probability* and the *Viterbi algorithm*.

The *posterior state probability* calculates the probability of being in a particular state at a specific time given the observation sequence and the model. These probabilities can then be used to find the state that most likely generated the observation for each time. Let $\gamma_t(i)$ be the probability of being in state S_i at time t given the observation sequence O and model λ, i.e.,

$$\gamma_t(i) = \Pr(q_t = S_i | O, \lambda), \quad 1 \leq i \leq N.$$

The posterior state probabilities can be calculated for all states and time points using the forward-backward probabilities

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}, \quad 1 \leq i \leq N, \quad 1 \leq t \leq T,$$

and we can determine the most probable state sequence by selecting the state at each time that maximizes $\gamma_t(i)$,

$$q_t = \text{argmax}_{1 \leq i \leq N}[\gamma_t(i)], \quad 1 \leq t \leq T. \tag{2.6}$$

A potential problem of using the posterior state probability to find the optimal state sequence is that the resultant sequence may be invalid if some of the state transitions are not permitted and therefore have a transition probability of zero. In order to avoid

such invalid sequences the Viterbi algorithm can be used to determine the optimal state sequence.

The *Viterbi algorithm* finds the single most likely state sequence that could have generated the observation sequence. It differs from the posterior state probability method because it chooses the state sequence that is globally optimum rather than individually selecting states at each position that are locally optimum. To compute the Viterbi algorithm, let $\delta_t(i)$ denote the highest log probability of the sequence of states that ends in state S_i at time t, then

$$\delta_t(i) = \max_{q_1,q_2,\cdots,q_{t-1}} \log[\Pr(q_1 q_2 \cdots q_t = S_i, o_1 o_2 \cdots o_t|\lambda)]. \tag{2.7}$$

Logarithms are used throughout the Viterbi calculations to avoid underflow. In order to retrieve the most likely state sequence, we need to keep track of the states giving rise to $\delta_i(i)$. Let $\psi_t(i)$ be the state at $t-1$ that produced the most probable state sequence ending in state S_i at time t. The Viterbi algorithm finds the most likely state sequence using the following procedure:

1. Initialization:
$$\delta_t(i) = \log(\pi_i) + \log[b_i(o_1)], \quad 1 \le i \le N$$
$$\psi_t(i) = 0.$$

2. Recursion:
$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i) + \log(a_{ij})] + \log[b_t(j)], \quad 2 \le t \le T \quad 1 \le j \le N$$
$$\psi_t(j) = \operatorname{argmax}_{1\le i\le N}[\delta_{t-1}(i) + \log(a_{ij})], \qquad 2 \le t \le T \quad 1 \le j \le N.$$

3. Termination:
$$\log P^* = \max_{1\le i\le N}[\delta_T(i)]$$
$$q_T^* = \operatorname{argmax}_{1\le i\le N}[\delta_T(i)].$$

4. State sequence backtracking:
$$q^* = \psi_{t+1}(q_{t+1}^*) \quad t = T-1, T-2, \ldots, 1.$$

Problem 3 The third problem of a HMM is concerned with estimating model parameters such that the probability of the observation sequence is maximized. This is generally solved using the Baum-Welch algorithm however we do not go into detail of the algorithm here as we do not make use of it. Instead we explain how we estimate parameters as they are introduced in our model.

2.2 XIBD

We implement a first order continuous time HMM to infer IBD between pairs of individuals using genotype data for SNPs, where we consider time here to be the genetic map distance (M) along a chromosome. Although the memoryless assumption of a Markov process is unlikely to hold for dense datasets in the presence of LD, McPeek and Sun[68] have shown it to be a good approximation and like Albrechtsen et al.[18] and Epstein et al.[69], we make this assumption also. Graphical representations of the model for different pairwise-ploidy combinations are shown in Figure 2.1 and we describe the model in more detail below.

2.2.1 State space

The state space in the model is the number of alleles shared IBD between a pair of individuals. A pair of diploid chromosomes (a pair of autosomes or the X chromosomes of two females), can either share 0, 1 or 2 alleles IBD. When the analysis includes at least one haploid chromosome (i.e., a male X chromosome) then at most 0 or 1 allele can be shared IBD. Therefore, the state space for a pair diploid chromosomes is $S = \{0, 1, 2\}$ while the state space when at least one chromosome is haploid is $S = \{0, 1\}$.

2.2.2 Initial probabilities

The probabilities associated with sharing 0, 1 or 2 alleles IBD are denoted ω_0, ω_1 and ω_2 respectively, and are used as the initial probabilities in our model. For known relationships, these probabilities can be calculated using identity coefficients. We use IdCoefs[70] for autosomes and have implemented the equivalent for the X chromosome (Appendix B). For example, a pair of siblings share 0, 1 and 2 alleles IBD on autosomes with probabilities $\omega_0 = 0.25$, $\omega_1 = 0.5$ and $\omega_2 = 0.25$. However if the relationships are unknown then these probabilities need to be estimated before we implement the HMM. We estimate the initial probabilities using the method of moments approach of Purcell et al.[41], described below.

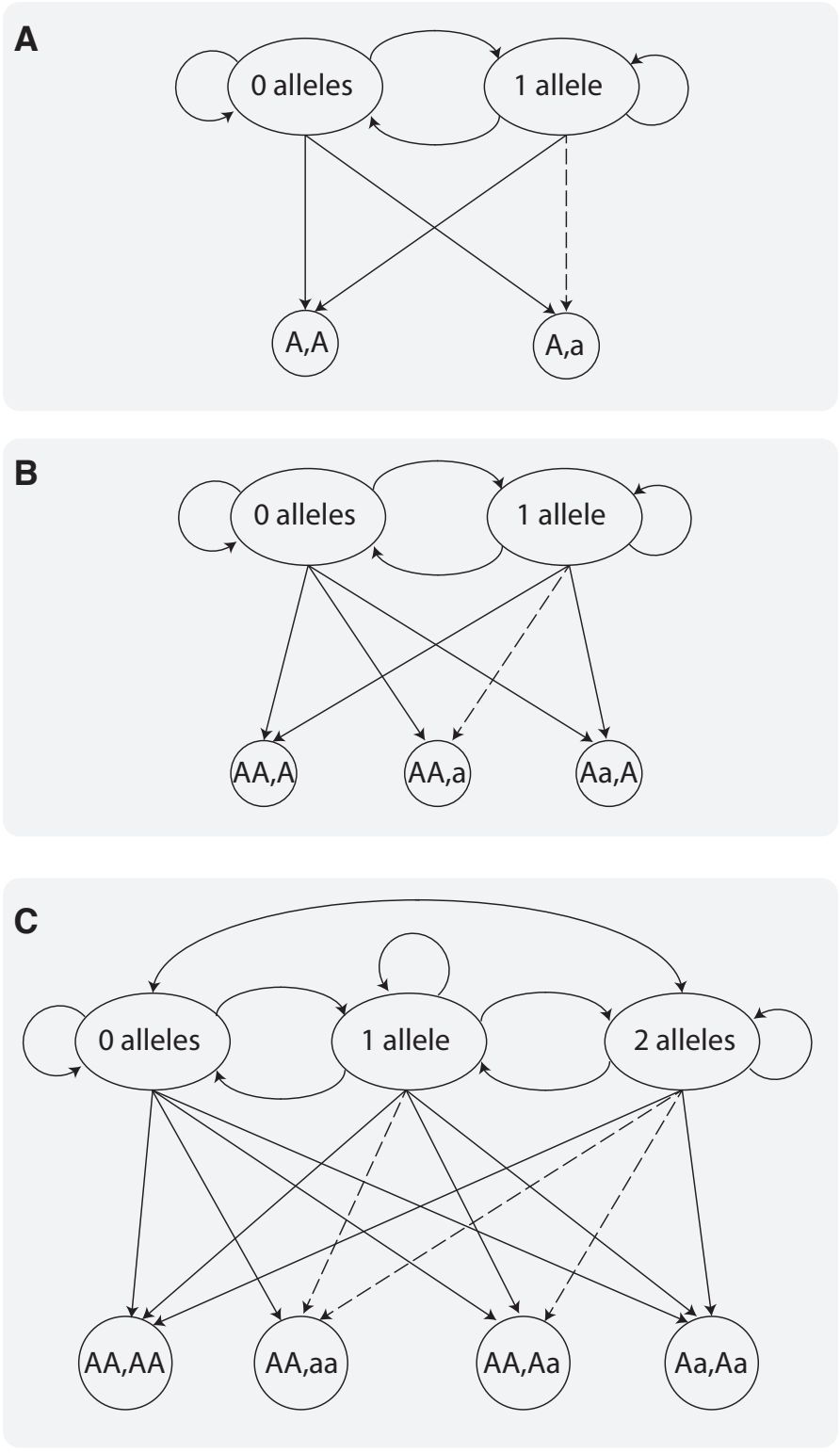


Figure 2.1: XIBD HMM for different pairwise-ploidy combinations. The states are the number of alleles shared IBD while the observations are genotypes for a pair of individuals. Dashed lines connect states to observations where the emission probability is zero. All other lines have non-zero probability distributions associated with them, calculated as in Tables 2.1-2.7. **A** The model used for pairs of haploid chromosomes. **B** The model used for pairs of chromosomes where one chromosome is haploid and the other chromosome is diploid. **C** The model used for pairs of diploid chromosomes.

Let $V, Z \in S$ denote IBS states and IBD states respectively. We wish to determine the probabilities of sharing 0, 1 and 2 alleles IBD, i.e.,

$$\omega_0 = \Pr(Z = 0),$$

$$\omega_1 = \Pr(Z = 1),$$

$$\omega_2 = \Pr(Z = 2).$$

The prior probability of IBS sharing can be expressed as a function of IBD sharing

$$\Pr(V = v) = \sum_{z=0}^{z=v} \Pr(V = v|Z = z) \Pr(Z = z). \tag{2.8}$$

where $\Pr(V = v|Z = z)$ can be expressed for each SNP in terms of population allele frequencies as in Table 2.1 while $\Pr(V = v)$ is either 1 or 0 depending if v matches the IBS state of the SNP. The allele frequencies can either be calculated from the input dataset if the individuals are of homogeneous population and the dataset is of sufficient size, otherwise a reference dataset such as HapMap (http://hapmap.ncbi.nlm.nih.gov/) can be used. These values can be averaged over all SNPs to obtain the observed and expected IBS values, after which equation 2.8 can be expanded and rearranged into

$$\Pr(Z = 0) = \sum_{t=1}^{T} \frac{\Pr(V = 0)}{\Pr(V = 0|Z = 0)},$$

$$\Pr(Z = 1) = \sum_{t=1}^{T} \frac{\Pr(V = 1) - \Pr(V = 0) \Pr(V = 1|Z = 0)}{\Pr(V = 1|Z = 1)},$$

$$\Pr(Z = 2) = \sum_{t=1}^{T} \frac{\Pr(V = 2) - \Pr(V = 1) \Pr(V = 2|Z = 1) - \Pr(V = 0) \Pr(V = 2|Z = 0)}{\Pr(V = 2|Z = 2)}.$$

Purcell et al.[41] constrain the estimates such that $\Pr(Z = z)$ are between 0 and 1, inclusive, and $\sum_{z=0}^{2} \Pr(Z = z) = 1$. Furthermore;

1. If $\Pr(Z = 0) > 1$ then $\Pr(Z = 0) = 1$, $\Pr(Z = 1) = 0$ and $\Pr(Z = 2) = 0$
2. If $\Pr(Z = 0) < 1$ then $\Pr(Z = 0) = 0$, $\Pr(Z = 1) = \Pr(Z = 1)/U$ and $\Pr(Z = 2) = \Pr(Z = 2)/U$ where $U = \Pr(Z = 1) + \Pr(Z = 2)$.

Table 2.1: Calculating $\Pr(V = v|Z = z)$ for a pair of individuals l and k; used when estimating the initial probabilities as described by Purcell et al. [41]. A diploid chromosome has ploidy = 2 while a haploid chromosome has ploidy = 1. The population allele frequency of the reference allele, A, is p_A and the allele frequency of the alternative allele, a, is $p_a = 1 - p_A$, where $p_A + p_a = 1$.

ploidy_l	ploidy_k	V	Z	$\Pr(V = v\|Z = z)$
1	1	0	0	$2p_Ap_a$
1	1	1	0	$p_A^2 + p_a^2$
1	1	0	1	0
1	1	1	1	1
2	1	0	0	$p_A^2q_a + p_Aq_a^2$
2	1	1	0	$p_A^3 + p_a^3 + 2p_A^2p_a + 2p_Ap_a^2$
2	1	0	1	0
2	1	1	1	1
2	2	0	0	$2p_A^2p_a^2$
2	2	1	0	$4p_A^3p_a + 4p_Ap_a^3$
2	2	2	0	$p_A^4 + p_a^4 + 4p_A^2p_a^2$
2	2	0	1	0
2	2	1	1	$2p_A^2p_a + 2p_Ap_a^2$
2	2	2	1	$p_A^3 + p_a^3 + p_A^2p_a + p_Ap_a^2$
2	2	0	2	0
2	2	1	2	0
2	2	2	2	1

3. If at least one chromosome is haploid then $\Pr(Z = 2) = 0$ and $\Pr(Z = 1) = 1 - \Pr(Z = 0)$.

2.2.3 Transition probability matrices

There are two state spaces in the model and as such we require two transition probability matrices. A transition probability matrix $A(t)$ can be computed by solving the Kolmogorov's forward equation

$$\begin{aligned} & & A'(t) &= A(t)Q \\ \text{subject to} & & A(0) &= I \end{aligned}$$

where Q is the transition rate matrix and t is the distance between adjacent markers in M. The solution of which is

$$A(t) = \exp(Qt).$$

Assuming the individuals are distantly related and there is no inbreeding, the transition rate matrix for a pair of diploid chromosomes is

$$Q = \begin{pmatrix} -\alpha\omega_1 & \alpha\omega_1 & 0 \\ \alpha\omega_0 & -\alpha(\omega_0 + \omega_2) & \alpha\omega_2 \\ 0 & \alpha\omega_1 & -\alpha\omega_1 \end{pmatrix},$$

and when at least one chromosome is haploid

$$Q = \begin{pmatrix} -\alpha\omega_1 & \alpha\omega_1 \\ \alpha\omega_0 & -\alpha\omega_0 \end{pmatrix},$$

where q_{ij} $(i \neq j)$ is the rate of departing from state i and arriving at state j and $\pi = (\omega_0, \omega_1, \omega_2)$ and $\pi = (\omega_0, \omega_1)$ are the stationary distributions, respectively. For example, the rate of departing from IBD = 0 to IBD = 1 for a pair of male X chromosomes is $\alpha\omega_1$. The parameter α controls the frequency of transitions between states and is a function of the number of meiosis m separating the pair of individuals and the recombination rate θ;

$$\alpha = -m \ln(1 - \theta).$$

The recombination rate is calculated from the Haldane map function [71]

$$\theta = \frac{1}{2}(1 - \exp^{-2t}),$$

while the number of meiosis is estimated according to Purcell et al. [41] and Albrechtsen et al. [18] as follows;

$$m = m_1 + m + 2, \qquad m_i = 1 - \log(x_i)/\log(2)$$

where

$$x_1 = \frac{\omega_1 + 2\omega_2 + \sqrt{(\omega_1 + 2\omega_2)^2 - 4\omega_2}}{2},$$

$$x_2 = \frac{\omega_2}{x_1}.$$

Here, m_1 and m_2 are the number of meioses from both lineages from a common ancestor for a pair of individuals.

To avoid computing the Taylor series for Q, which is often difficult, we can transform

Q into Jordan canonical form

$$Q = CJC^{-1}$$

and solve for the simpler expression

$$A(t) = \exp(Qt) = C\exp(Jt)C^{-1}. \tag{2.9}$$

For the simplest scenario, when at least one chromosome is haploid, we get

$$C = \begin{pmatrix} 1 & -\alpha\omega_1 \\ 1 & -\alpha\omega_0 \end{pmatrix} \qquad J = \begin{pmatrix} 0 & 0 \\ 0 & -\alpha \end{pmatrix},$$

and after some algebraic manipulation and substitution we get the following transition probability matrix

$$A(t) = \begin{pmatrix} \omega_0 + \omega_1\exp(-\alpha t) & \omega_1(1-\exp(-\alpha t)) \\ \omega_0(1-\exp(-\alpha t)) & \omega_1 + \omega_0\exp(-\alpha t) \end{pmatrix}.$$

The transition matrix for a pair of diploid chromosomes, as from Albrechtsen et al. [18], is given by

$$\begin{pmatrix} 1-(1-\exp(-\alpha t))\omega_1 - T_{0,2} & (1-\exp(-\alpha t))\omega_1 & T_{0,2} \\ (1-\exp(-\alpha t))\omega_0 & (1-\exp(-\alpha t))\omega_1 + \exp(-\alpha t) & (1-\exp(-\alpha t))\omega_2 \\ T_{2,0} & (1-\exp(-\alpha t))\omega_1 & 1-(1-\exp(-\alpha t))\omega_1 - T_{2,0} \end{pmatrix}$$

where

$$T_{0,2} = \frac{\exp(-\alpha\omega_1 t)\omega_2}{\omega_1 - 1} + \exp(-\alpha t)\omega_1 + \frac{\exp(-\alpha t)\omega_0\omega_1}{\omega_1 - 1} + \omega_2$$

and

$$T_{2,0} = \frac{\exp(-\alpha\omega_1 t)\omega_0}{\omega_1 - 1} + \exp(-\alpha t)\omega_1 + \frac{\exp(-\alpha t)\omega_2\omega_1}{\omega_1 - 1} + \omega_0.$$

$T_{i,j}$ is the probability of transitioning from state i to j.

2.2.4 Emission probabilities

Our methodology detects IBD between pairs of individuals using genotype data for biallelic SNPs. Therefore the observations in the model are pairs of genotypes, where the pairwise

combination of genotypes depends on the ploidy of the chromosomes under consideration. A haploid chromosome can either have the reference allele or the alternative allele at a SNP, producing the genotypic set $G = \{A, a\}$. In contrast, a diploid chromosome can either be homozygous reference, heterozygous or homozygous alternative at a SNP, giving the genotypic set $G = \{AA, Aa, aa\}$. This results in three pairwise combinations of observation sets, Φ^2;

$$\{\{A, a\} \times \{A, a\}\},$$

$$\{\{A, a\} \times \{AA, Aa, aa\}\},$$

$$\{\{AA, Aa, aa\} \times \{AA, Aa, aa\}\}.$$

It follows that there are three sets of emission probabilities (Table 2.2), one for each observation set, which are functions of the chromosome ploidies (dependent on the individual's genders), the state space, the observed genotype pair, and the population allele frequencies. Emission probabilities can be extended to account for genotyping errors, missing data and LD as follows.

Table 2.2: Emission probabilities for a pair of individuals. A diploid chromosome has ploidy = 2 while a haploid chromosome has ploidy = 1. By symmetry $\Pr(G_i^{l,k}|Z_i = z) = \Pr(G_i^{k,l}|Z_i = z)$.

ploidy$_l$	ploidy$_k$	G_i^l	G_i^k	$Z_i = 0$	$Z_i = 1$	$Z_i = 2$
1	1	A	A	p_A^2	p_A	0
1	1	A	a	$2p_Ap_a$	0	0
2	1	AA	A	p_A^3	p_A^2	0
2	1	AA	a	$p_A^2p_a$	0	0
2	1	Aa	A	$2p_A^2p_a$	p_Ap_a	0
2	2	AA	AA	p_A^4	p_A^3	p_A^2
2	2	AA	aa	$2p_A^2p_a^2$	0	0
2	2	AA	Aa	$4p_A^3p_a$	$2p_A^2p_a$	0
2	2	Aa	Aa	$4p_A^2p_a^2$	$p_A^2p_a + p_Ap_a^2$	$2p_Ap_a$

Genotyping errors

Most genetic datasets will contain a small number of genotyping errors, which can result in incorrect IBD inference and/or reduced performance[39]. As such, we account for genotyping errors in our model. Let ϵ denote the error rate, G_i^g denote the observed genotype

of individual g at SNP i and let $G_i^{g'}$ denote the true genotype individual g at SNP i. The probabilities of the observed genotypes given the true genotypes are in Tables 2.3 and 2.4 for haploid and diploid chromosomes respectively. Like Albrechtsen et al.[18], these probabilities are included in the calculation of emission probabilities as follows, where ϵ is assumed known prior to analysis.

$$\Pr(G_i^{l,k}|Z_i = z, \epsilon) = \sum_{G_i^{l',k'} \in \Phi^2} \Pr(G_i^{l',k'}|Z_i = z)(\prod_{g \in \{l,k\}} \Pr(G_i^g|G^{g'}, \epsilon)) \tag{2.10}$$

Table 2.3: Genotyping error probabilities of the observed genotype G_i^l given the true genotype $G_i^{l'}$ and an error rate of ϵ for a haploid chromosome.

$P(G_i^l\|G_i^{l'}, \epsilon)$	$G_i^l = A$	$G_i^l = a$
$G_i^{l'} = A$	1-ϵ	ϵ
$G_i^{l'} = a$	ϵ	1-ϵ

Table 2.4: Genotyping error probabilities of the observed genotype G_i^l given the true genotype $G_i^{l'}$ and an error rate of ϵ for a diploid chromosome.

$P(G_i^l\|G_i^{l'}, \epsilon)$	$G_i^l = A$	$G_i^l = Aa$	$G_i^l = AA$
$G_i^{l'} = A$	$(1-\epsilon)^2$	$2(1-\epsilon)\epsilon$	ϵ^2
$G_i^{l'} = Aa$	$(1-\epsilon)\epsilon$	$(1-\epsilon)^2 + \epsilon^2$	$(1-\epsilon)\epsilon$
$G_i^{l'} = AA$	ϵ^2	1-ϵ	$(1-\epsilon)^2$

Missing data

SNPs with missing genotypes are sometimes removed from analyses, resulting in reduced datasets and a loss of information. To avoid this we allow for missing data by calculating the summation of the emission probabilities over the genotypic observation sets for SNPs with missing genotypes.

Linkage disequilibrium

The assumption of a memoryless Markov chain is unlikely to hold in the presence of LD. To overcome this, Purcell et al.[41] recommends pruning SNPs such that the remaining SNPs are in approximate LE. However, reducing datasets in such a way can result in the loss of potentially informative SNPs and in some instances less than 5% of the original

data may remain. Alternatively, one could account for LD in the model. Albrechtsen et al.[18] account for LD through conditional emission probabilities, where the genotype of SNP i is conditioned on by the genotype of a single SNP h amongst the previous s SNPs that is in the highest LD with SNP i.

Accounting for LD by conditioning on even a single SNP can add considerable time to the already computationally inefficient model. Additionally, single SNP conditioning may not account for all LD in a dataset, especially dense SNP datasets, which can result in detection of unwanted background sharing. As such, we implement two models to accommodate for LD.

1. Like Purcell et al.[41], model 1 assumes the SNPs are in LE, which typically requires thinning of datasets prior to use. However, datasets with dense SNPs in LD can be used at the expense of false IBD segments being reported[17]

2. Like Albrechtsen et al.[18], LD is implicitly accounted for in model 2 using conditional emission probabilities (Equation 2.11). This required calculation of joint genotype probabilities, which are provided in Tables 2.5, 2.6 and 2.7. Haplotype frequencies are calculated from genotype data as described in Clayton and Leung[72]. While it may be desirable to calculate haplotype frequencies from phased data rather than unphased data, we acknowledge that phased data is not always available and phasing can be time consuming to perform. Furthermore, large datasets are typically required for phase inference.

Unlike Purcell et al.[41] and Albrechtsen et al.[18], reference datasets are provided with XIBD. These datasets are the combined HapMap Phase II and III genotypes and allele frequencies from build 19[73]; allowing the user to choose between the 11 HapMap population. Furthermore, given a homogeneous population, we allow the user to calculate the necessary frequencies from the input dataset itself or to specify their own homogeneous reference dataset of matching population.

2.3 Summary

Here we introduced a model for IBD inference of both haploid and diploid chromosomes, namely XIBD. XIBD allows for pairwise analysis of the X chromosome without the need

to duplicate the male X chromosome. Furthermore, the model can be used for analysis of non-human organisms with haploid (or diploid) genomes.

Simply, XIBD implements a HMM to infer the numbers of alleles shared IBD between a pair of individuals using SNP genotype data. With advances in technologies producing datasets with increasing SNP density, it is important to account for both genotyping errors and LD when developing a model. We account for genotyping errors through the inclusion of an error term in the calculation of the emission probabilities, while LD is accounted for by either thinning the dataset prior to analysis to produce a collection of SNPs in approximate LE, or implementing emission probabilities that condition on the genotypes of a pervious SNP. The model presented here fills a void in what is missing from IBD methodologies and allows for a more complete analysis of the human genome and the genomes of other organisms.

$$\Pr(G_i^{l,k}|G_h^{l,k}, Z_i = Z_h = z, \epsilon) = \frac{\Pr(G_i^{l,k}, G_h^{l,k}|Z_i = Z_h = z, \epsilon)}{\Pr(G_h^{l,k}|Z_h = z, \epsilon)} \tag{2.11}$$

$$= \frac{\sum_{G_i^{l',k'}, G_h^{l',k'} \in \Phi^4} \Pr(G_i^{l',k'}, G_h^{l',k'}|Z_i = Z_h = z)(\prod_{g \in \{l,k\}} \Pr(G_i^g|G_i^{g'}, \epsilon))(\prod_{g \in \{l,k\}} \Pr(G_h^g|G_h^{g'}, \epsilon))}{\sum_{G_h^{l',k'} \in \Phi^2} \Pr(G_h^{l',k'}|Z_h = z)(\prod_{g \in \{l,k\}} \Pr(G_i^g|G^{g'}, \epsilon))}$$

Table 2.5: Joint probabilities for observing two male X chromosome genotypes at positions i and h

G_i^l	G_i^k	G_h^l	G_h^k	$Z_i = 0$	$Z_i = 1$
B	B	A	A	p_{BA}^2	p_{BA}
B	B	A	a	$2p_{BA}p_{Ba}$	0
B	b	A	A	$2p_{BA}p_{bA}$	0
B	b	A	a	$2p_{BA}p_{ba} + 2p_{Ba}p_{bA}$	0

Table 2.6: Joint probabilities for observing a female/male pair of X chromosome genotypes at positions i and h

G_i^l	G_i^k	G_h^l	G_h^k	$Z_i = 0$	$Z_i = 1$
BB	B	AA	A	p_{BA}^3	p_{BA}^2
BB	B	AA	a	$p_{BA}^2 p_{Ba}$	0
BB	B	Aa	A	$2p_{BA}^2 p_{Ba}$	$2p_{BA}p_{Ba}$
BB	b	AA	A	$p_{BA}^2 p_{bA}$	0
BB	b	AA	a	$p_{BA}^2 p_{ba}$	0
BB	b	Aa	A	$2p_{BA}p_{Ba}p_{bA}$	0
Bb	B	AA	A	$2p_{BA}^2 p_{bA}$	$2p_{BA}p_{bA}$
Bb	B	AA	a	$2p_{BA}p_{bA}p_{Ba}$	0
Bb	B	Aa	A	$2p_{BA}^2 p_{ba} + 2p_{Ba}p_{bA}p_{BA}$	$p_{BA}p_{ba}$

Table 2.7: Joint probabilities for observing a pair of haploid chromosome genotypes at positions i and h

G_i^l	G_i^k	G_h^l	G_h^k	$Z_i = 0$	$Z_i = 1$	$Z_i = 2$
BB	BB	AA	AA	p_{BA}^4	p_{BA}^3	p_{BA}^2
BB	BB	AA	Aa	$4p_{BA}^3 p_{Ba}$	$2p_{BA}^2 p_{Ba}$	0
BB	BB	AA	aa	$2p_{BA}^2 p_{Ba}^2$	0	0
BB	BB	Aa	Aa	$4p_{BA}^2 p_{Ba}^2$	$p_{BA}^2 p_{Ba} + p_{Ba}^2 p_{BA}$	$2p_{BA}p_{Ba}$
BB	Bb	AA	AA	$4p_{BA}^3 p_{bA}$	$2p_{BA}^2 p_{bA}$	0
BB	Bb	AA	Aa	$4p_{BA}^2 p_{bA} p_{Ba} + 4p_{BA}^3 p_{ba} + 8p_{BA}^2 p_{bA} p_{Ba}$	$2p_{BA}^2 p_{ba} + 2_{BA} p_{Ba} p_{bA}$	0
BB	Bb	AA	aa	$4p_{BA}^2 p_{ba} p_{Ba} + 4p_{Ba}^2 p_{bA} p_{BA}$	0	0
BB	Bb	Aa	Aa	$8p_{BA}^2 p_{ba} p_{Ba} + 8p_{BA} p_{Ba}^2 p_{bA}$	$2p_{BA} p_{ba} p_{Ba} + 2p_{BA} p_{Ba} p_{bA}$	0
BB	bb	AA	AA	$2p_{BA}^2 p_{bA}^2$	0	0
BB	bb	AA	Aa	$4p_{BA} p_{Ba} p_{bA}^2 + 4p_{BA}^2 p_{bA} p_{ba}$	0	0
BB	bb	AA	aa	$2p_{BA}^2 p_{ba}^2 + 2p_{Ba}^2 p_{bA}^2$	0	0
BB	bb	Aa	Aa	$8p_{ba} p_{bA} p_{BA} p_{Ba}$	0	0
Bb	Bb	AA	AA	$4p_{BA}^2 p_{bA}$	$p_{BA} p_{bA}^2 + p_{BA}^2 p_{bA}$	$2p_{BA}p_{bA}$
Bb	Bb	AA	Aa	$8p_{BA}^2 p_{bA} p_{ba} + 8p_{BA} p_{bA}^2 p_{Ba}$	$2p_{BA} p_{bA} p_{ba} + 2p_{BA} p_{bA} p_{Ba}$	0
Bb	Bb	AA	aa	$8p_{ba} p_{bA} p_{BA} p_{Ba}$	0	0
Bb	Bb	Aa	Aa	$4p_{BA}^2 p_{ba}^2 + 8p_{ba} p_{bA} p_{BA} p_{Ba} + 4p_{Ba}^2 p_{bA}^2$	$p_{BA}^2 p_{ba} + p_{BA} p_{ba}^2 + p_{Ba}^2 p_{bA} + p_{Ba} p_{bA}^2$	$2p_{BA}p_{ba} + 2p_{Ba}p_{bA}$

Chapter 3

Simulation studies and software description for XIBD

We performed simulation studies to assess the power and accuracy of XIBD in determining IBD segments of varying lengths in the presence of varying levels of LD. This study was designed to assist with model and SNP selection for use with XIBD. We also show a surprising difference in power and accuracy between pairs of chromosomes with different ploidy combinations. Furthermore we perform comparisons with other popular IBD methodologies, GERMLINE[47] and fastIBD[49], and provide a critical summary of the efficiency of XIBD.

3.1 Simulating artificial IBD segments

We simulated SNP genotype data for the X chromosome for pairs of individuals separated from 1 to 25 generations (siblings to 24th cousins). All generation were simulated using all female lineages such that individuals separated by 25 generations were likely to have on average IBD segments of length 2cM, which is the smallest length detected with high power by many IBD algorithms. For each of the 25 generations we simulated 1,000 diploid pairs of related chromosomes (mimicking a pair of female's X chromosomes or a pair of autosomes), 1,000 diploid/haploid pairs of chromosomes (mimicking a pair of X chromosomes belonging to one female and one male) and 1,000 haploid pairs of chromosomes (mimicking a pair of male's X chromosomes). We selected SNPs to be included in the analysis from the Illumina HumanOmni2.5 platform. This platform contains 16,382 X chromosome SNPs for which HapMap Phase II data was available for the CEU population. This reflects a

dataset of SNPs that could typically be extracted using LINKDATAGEN[74] from either SNP or NGS data. All SNPs with minor allele frequencies (MAF) less than 1% were excluded from analyses as these SNPs are not considered common to the population and are likely to result in false positive IBD calls due to extremely distant relatedness, which we are not interested in identifying. 13,509 SNPs remain following minor allele frequency filtering, corresponding to approximately 1 SNP per 0.01 cM.

Data was simulated to reflect X chromosome inheritance patterns using pedigree information as follows. Given a pedigree with all female lineage, haplotypes were generated for all male founders using algorithm 1. This algorithm was repeated twice to generate two haplotypes for female founders. Simulating haplotypes in such a way provides knowledge of haplotype phase, which is not necessary for XIBD, however can improve the performance of GERMLINE. Algorithm 1 requires population allele frequencies and haplotype frequencies, which were calculated using the HapMap phase II CEU genotype data. Haplotype frequencies between adjacent SNPs were calculated as described in Clayton and Leung[72]. While it may be desirable to calculate haplotype frequencies from phased data, this is time consuming to perform and frequency inference using Clayton and Leung for pairs of SNPs is sufficient for our purposes. Following founder haplotype simulation, recombination could be used to generate haplotypes for all non-founders in the pedigree according to algorithm 2. Here we assume that recombination follows an exponential distribution with mean 1 Morgan and that recombination only occurs between a female's two X chromosomes. All non-founders inherit a mosaic of their X chromosomes, while female non-founders also inherit their father's non-recombined X chromosome. Data was simulated to ensure that each pair of individuals shared at least one segment of IBD and only female siblings could share regions with two alleles IBD. Table 3.1 gives the number of IBD segments of certain lengths simulated for different ploidy combinations. Unphased genotype calls were generated from the phased haplotype data for use with XIBD and fastIBD.

3.1.1 Estimating power, accuracy, under- and overestimation of IBD

We use the same definitions of power, accuracy, under- and overestimation of IBD segments as defined in Browning and Browning[52]. We define power as the average proportion of a segment that is detected as a function of the size of the true IBD segment in cM; where undetected segments are included in this calculation. Accuracy is calculated as the

Algorithm 1 Simulating a haplotype

let N equal the total number of SNPs;
let p_i denote the population allele frequency at SNP i;
let $p_{i,i-1}$ denote the conditional probability of allele A at SNP_i given the allele at $\text{SNP}_i - 1$;
for SNP = 1 **do**
    generate $y \sim$ Unif(0,1)
    if $y < p_i$ **then**
        allele A is chosen for SNP = 1
    else
        allele B is chosen for SNP = 1
    end if
end for
for SNP $i = 2$ to N **do**
    calculate $p_{i,i-1}$
    generate $y \sim$ Unif(0,1)
    if $y < p_{i,i-1}$ **then**
        allele A is chosen for SNP_i
    else
        allele B is chosen for SNP_i
    end if
end for

probability that at least 50% of a detected segment is true as a function of the reported size of the detected segment. Underestimation is defined as the average length of an IBD segment that was not detected as a function of the size of the true IBD segment in cM, conditional on at least part of the segment being detected. This includes multiple gaps throughout a segment as well as underestimation of segment ends. Finally, overestimation of IBD is calculated as the average amount that a true IBD segment was overestimated by, conditional on at least part of the segment being detected. In instances where a detected IBD segment overlaps multiple true IBD segments, overestimation of one IBD segment is measured until the nearest IBD segment that was also detected by the same segment.

3.2 Evaluating XIBD by varying LD with different ploidies

We performed an analysis to assess the power and accuracy of each model under varying levels of LD. We pruned SNPs according to pairwise SNP correlations (R^2) greater than 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, and 1, where $R^2 > 0.3$ removes all SNPs with pairwise correlations greater than 0.3 while $R^2 > 1$ does not remove any SNPs in LD. SNP correlations were calculated from genotype allele counts for all individuals using PLINK[41]. Table 3.2 displays the number of SNPs remaining after LD filtering was

Algorithm 2 Simulating recombination

let $X_L = 1.8$ denote the length of the X chromosome in Morgans;
let X_1 and X_2 denote two homologous chromosomes;
generate $y \sim$ Unif(0,1);
if $y < 0.5$ **then**
    X1 is chosen as the start chromosome;
else
    X2 is chosen as the start chromosome;
end if

generate $z \sim$ Exp(1);
$t = z$;
if $t < X_L$ **then**
    recombination occurs at t;
else
    recombination does not occur and the offspring inherits a non-recombined chromosome;
end if

while $t < X_L$ **do**
    generate $z \sim$ Exp(1);
    $t = t + z$
    if $t < X_L$ **then**
        recombination at t;
    else
        stop, no more recombination;
    end if
end while

Table 3.1: The number of IBD segments simulated for various segment lengths (cM)

IBD segment length (cM)	diploid pairs	diploid/haploid pairs	haploid/haploid pairs
(0, 0.5]	4,839	4,761	4,579
(0.5, 1]	3,488	3,358	2,923
(1, 1.5]	2,433	2,710	3,112
(1.5, 2]	2,198	2,221	2,335
(2, 3]	3,130	3,267	3,146
(3, 4]	1,961	1,707	1,941
(4, 5]	1,391	1,414	1,434
(5, 10]	3,377	3,423	3,481
> 10	7,030	6,001	6,214

Table 3.2: The number of SNPs remaining once SNPs in LD are removed. Pairs of SNPs were removed if their pairwise correlation (R^2) based on genotype allele counts was more than the specified threshold. Setting $R^2 = 1.00$ will not remove any SNPs.

R^2	SNPs remaining
0.30	7,906
0.40	9,267
0.50	10,397
0.60	11,236
0.70	11,996
0.80	12,973
0.90	13,395
0.95	13,442
0.99	13,450
1.00	13,509

performed. Figures 3.1, 3.2 and 3.3 each display the results for model 1 and model 2 for pairs of diploid chromosomes, pairs of chromosomes where one chromosome is diploid and the other chromosome is haploid, and pairs of haploid chromosomes, respectively.

For all ploidy combinations, model 1 (pruning) has similar power to detect IBD segments of all lengths as model 2 (conditioning), however IBD segments less than 2cM are more likely to be true positives when detected by model 2. This suggests that greater caution should be taken when interpreting the results from model 1 if stringent post-analysis filtering of small IBD segments has not been performed. In contrast, if an IBD segment of less than 2cM is detected with model 2, there is at least a 50% chance that it will be real. Accounting for LD in the algorithm, even if not complete, results in better performance than not accounting for it at all.

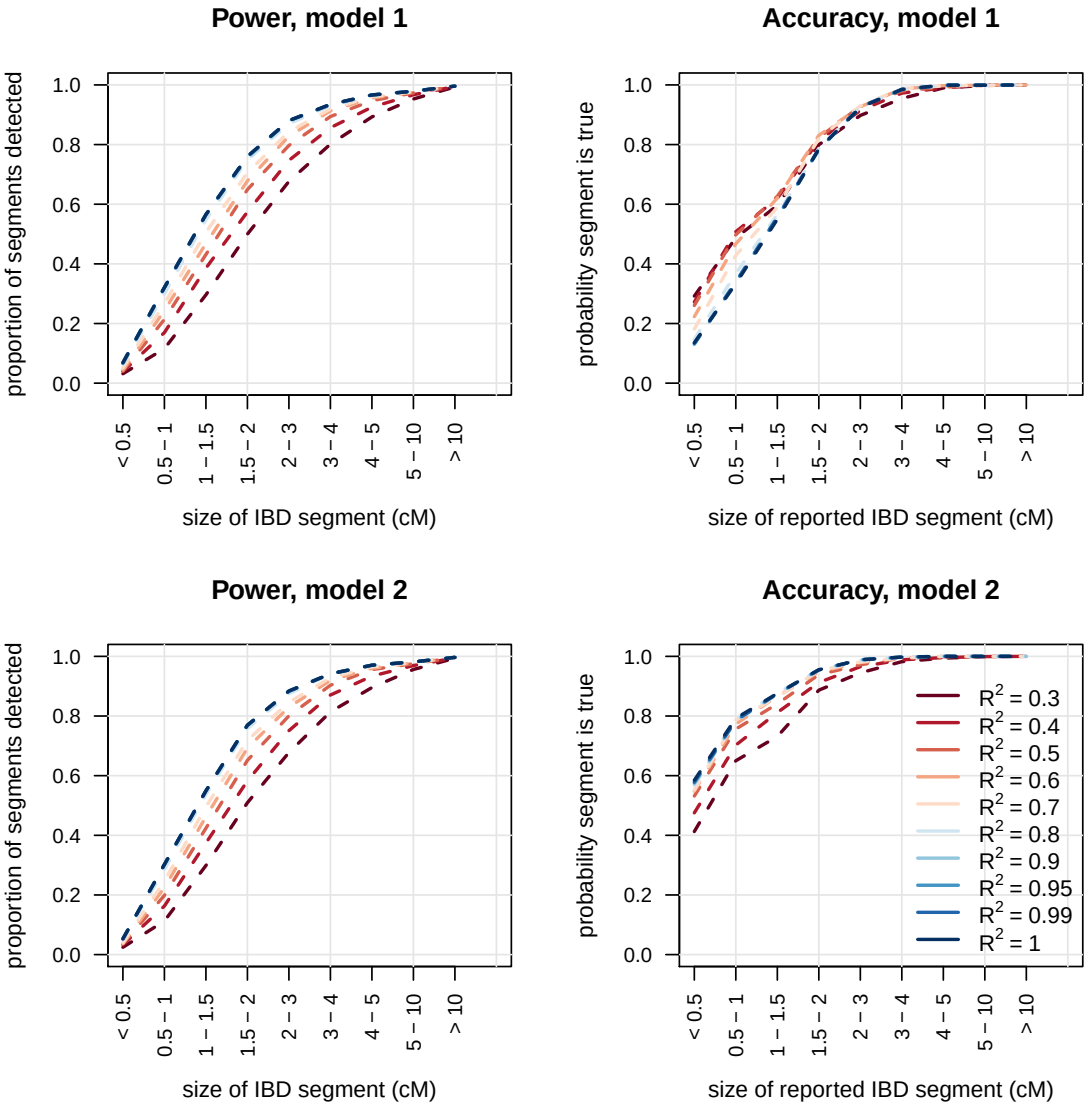


Figure 3.1: Power and accuracy results for XIBD between 2 diploid chromosomes, calculated across various segment sizes and LD filtering levels.

We also find that pruning SNPs using lower thresholds of LD (i.e. $R^2 = 0.3$) reduces the power of XIBD to detect segments of many sizes, using both models, while the accuracy of model 1 (pruning) is highest for heavily pruned datasets (i.e. $R^2 = 0.3$) whereas the accuracy of model 2 (conditioning) is highest for least pruned datasets (i.e. $R^2 = 1$). This difference is likely attributed to the choice of SNP that is conditioned on when using model 2, which may be poorly selected when datasets are heavily filtered according to LD thresholds. Additionally, there is little difference in XIBD performance when SNPs are pruned according to $R^2 \geq 0.8$. This suggests that it is beneficial to keep more SNPs in the

analysis, to an extent, at the risk of detecting slightly more false positive IBD segments of smaller sizes.

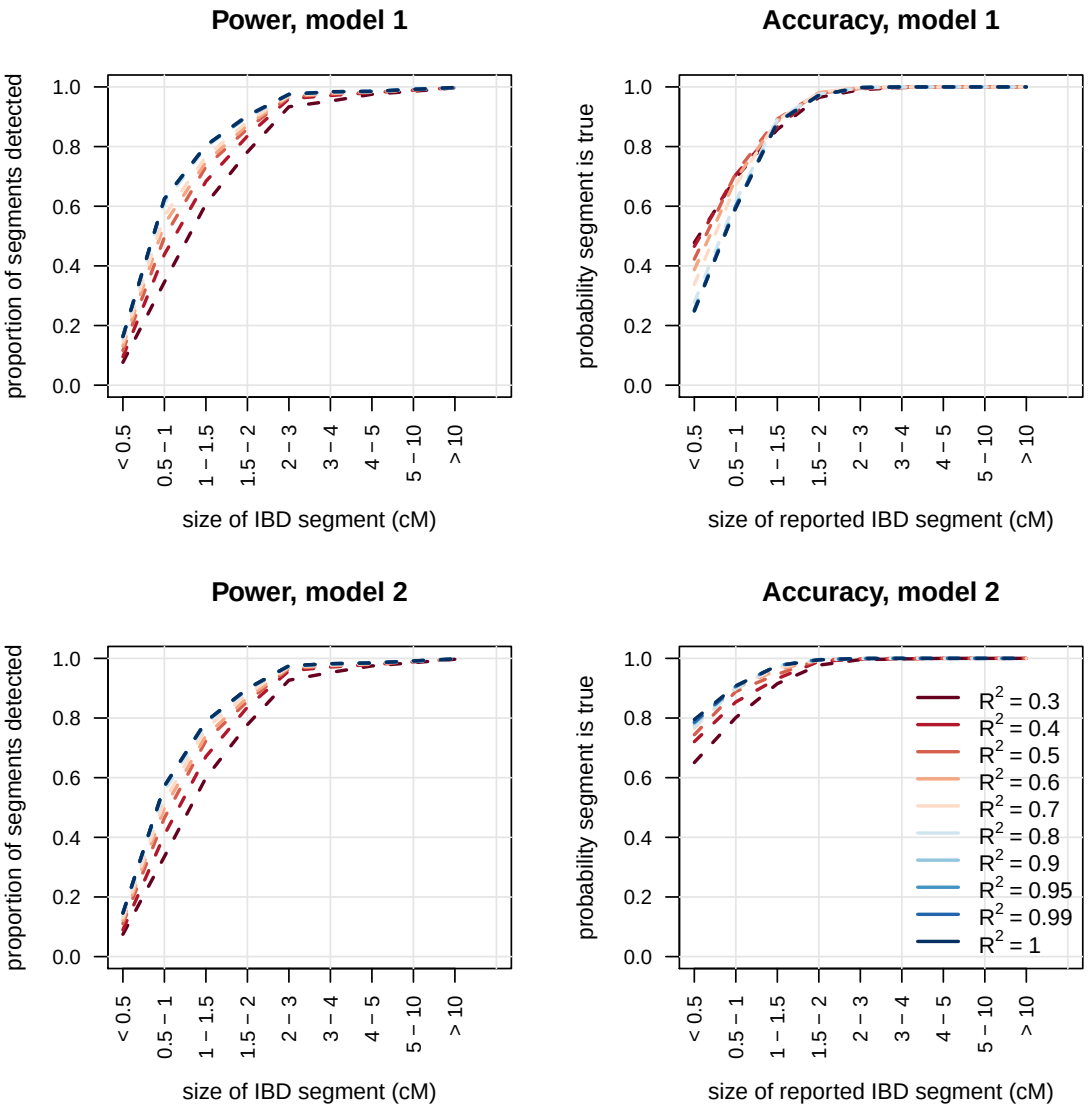


Figure 3.2: Power and accuracy results for XIBD between one diploid and one haploid chromosome, calculated across various segment sizes and LD filtering levels.

Lastly, there is a considerable difference in the performance of XIBD for different ploidy combinations (Figure 3.4). In particular, XIBD performs exceptionally well when applied to pairs of haploid chromosomes, while performance decreases as diploid chromosomes are included in the analysis. This result is not surprising as the state space and probability distributions are simplified for haploid chromosomes and IBD is more obvious as haplotype phase is known.

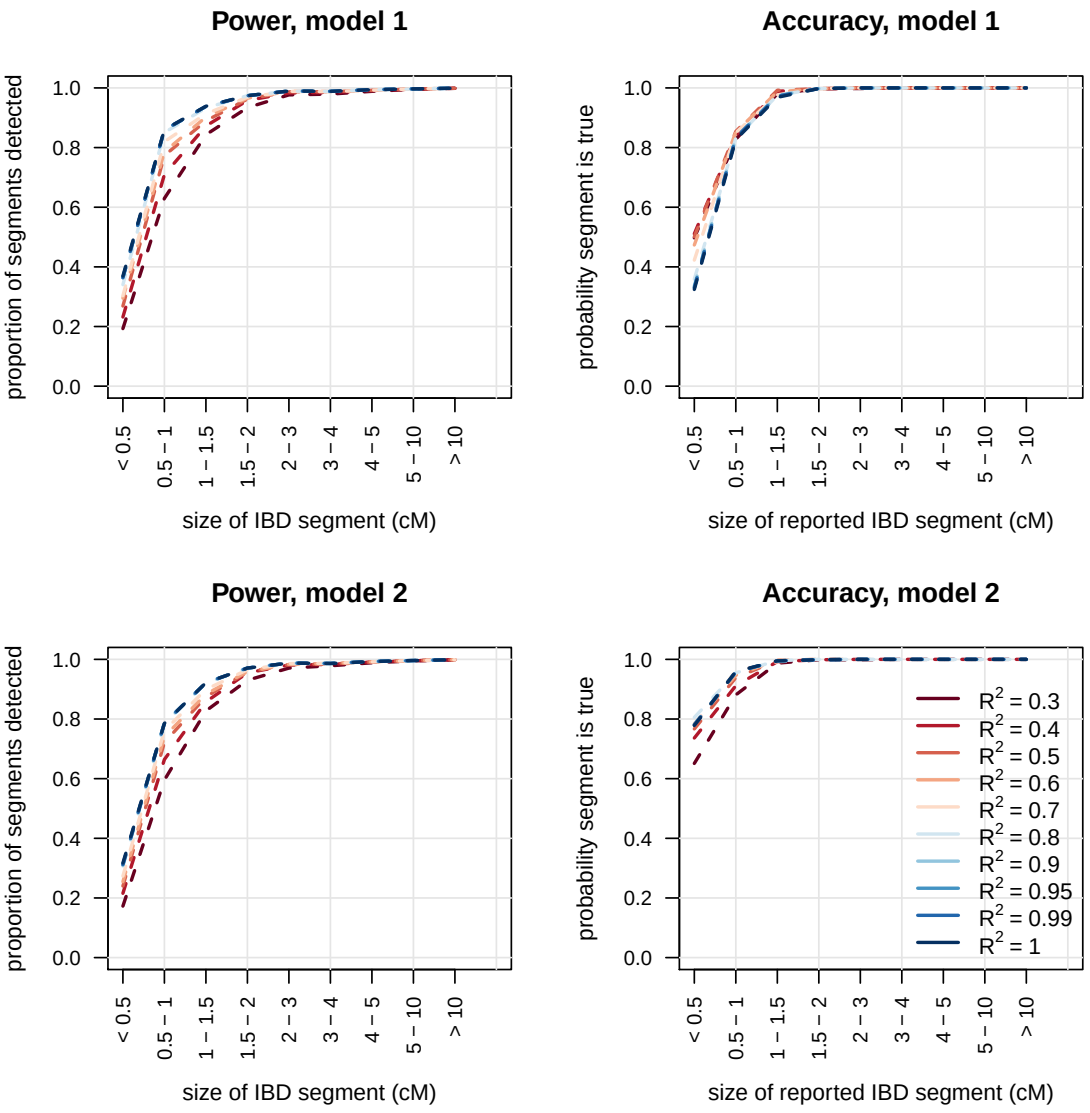


Figure 3.3: Power and accuracy results for XIBD between 2 haploid chromosomes, calculated across various segment sizes and LD filtering levels.

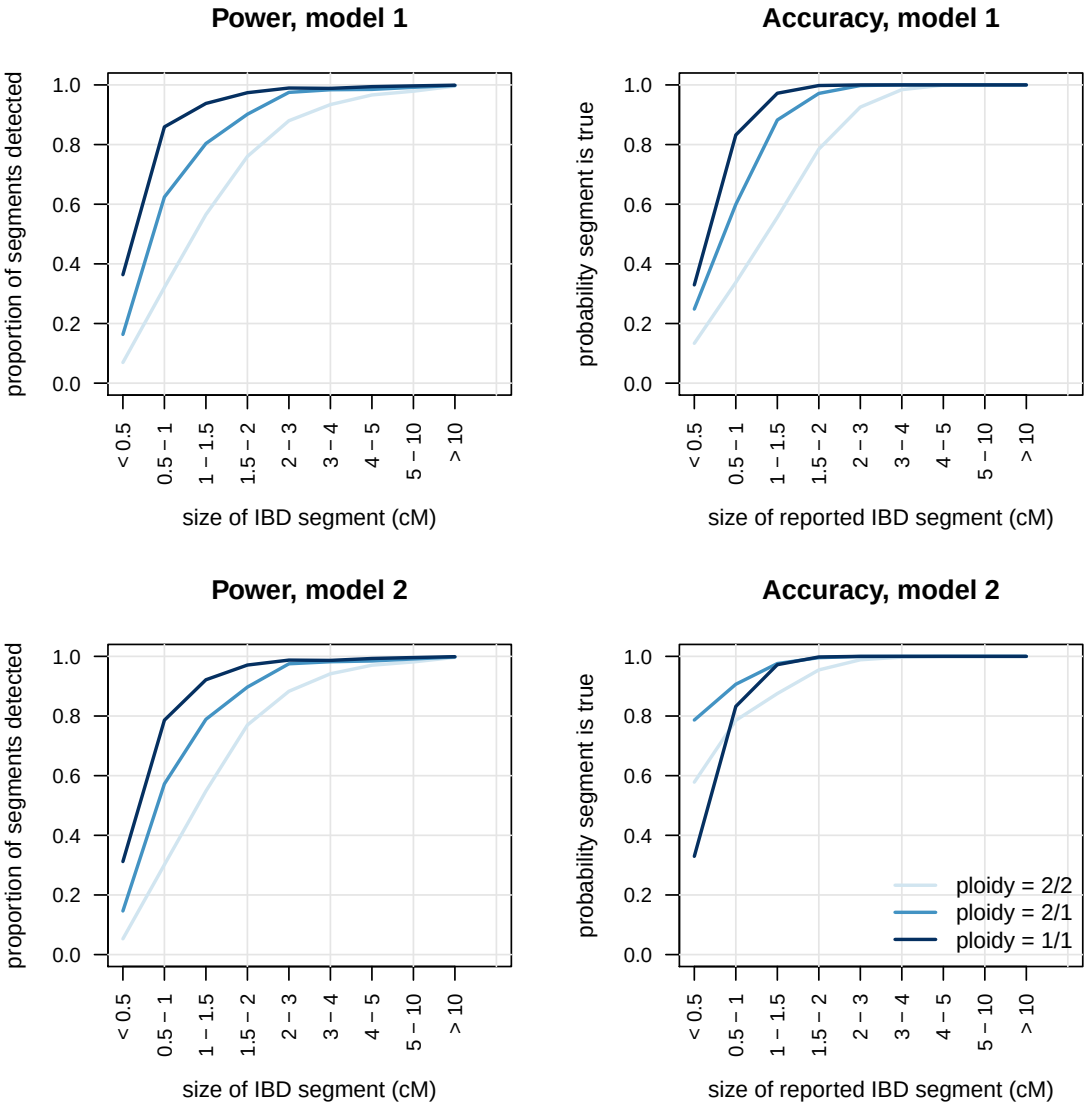


Figure 3.4: Power and accuracy results for XIBD across ploidies, calculated across various segment sizes with $R^2 = 0.99$. Ploidy = 2/2 denotes a pair of diploid chromosomes, ploidy = 2/1 denotes a pair of chromosomes where one chromosome is diploid and the other chromosome is haploid, and ploidy = 1/1 denotes a pair of haploid chromosomes.

3.3 Comparison of XIBD, GERMLINE and fastIBD

3.3.1 Model parameters

We compared the performance of XIBD to GERMLINE and fastIBD using the simulated data described earlier. Specifically, the dataset containing 13,509 SNP was used, where SNPs with MAF $< 1\%$ were removed and no LD filtering was performed. GERMLINE can be implemented on haploid chromosomes, however the same cannot be said for fastIBD. We chose to assess the performance of fastIBD, even though haploid chromosomes are not properly accounted for in the probability distributions, as the localized haplotype cluster model that defines the BEAGLE HMM[61], which is the basis of fastIBD, is used by all BEAGLE IBD algorithms and thus is widely used in IBD tools.

GERMLINE version 1.5.1 was implemented with parameters "-bits 32 -min_m 0.5 -err_hom 1 -err_het 1 -haploid", as in in Gusev et al.[75] with the exception of "-min_m". The parameter "-min_m" defines the minimum reported length of an IBD segment in cM. We selected this parameter to be 0.5 cM (opposed to 1 cM as in Gusev et al.[75]) to determine IBD performance of smaller segment lengths. The parameter "-haploid" makes use of haplotype phase, and although XIBD performs analyses on unphased data, we assessed GERMLINE using phased data for optimal performance.

We performed analyses of fastIBD using version 3.3.2. We performed ten runs of fastIBD using different random number seeds and merged the results as recommended[49]. Within each run we performed ten iterations of the phasing algorithm and used default values for all other parameters. The dataset was large for ten iterations to be performed, so we created subsets of 500 individuals to perform analyses on and merged results. Each subset contained 10 pairs of individuals from each generation from siblings to 24th cousins, where each individual shared IBD with only one other individual per subset. We chose subsets of 500 individuals as this has been shown to achieve high power and accuracy with fastIBD[52]. The results are summarised across all subsets for all runs.

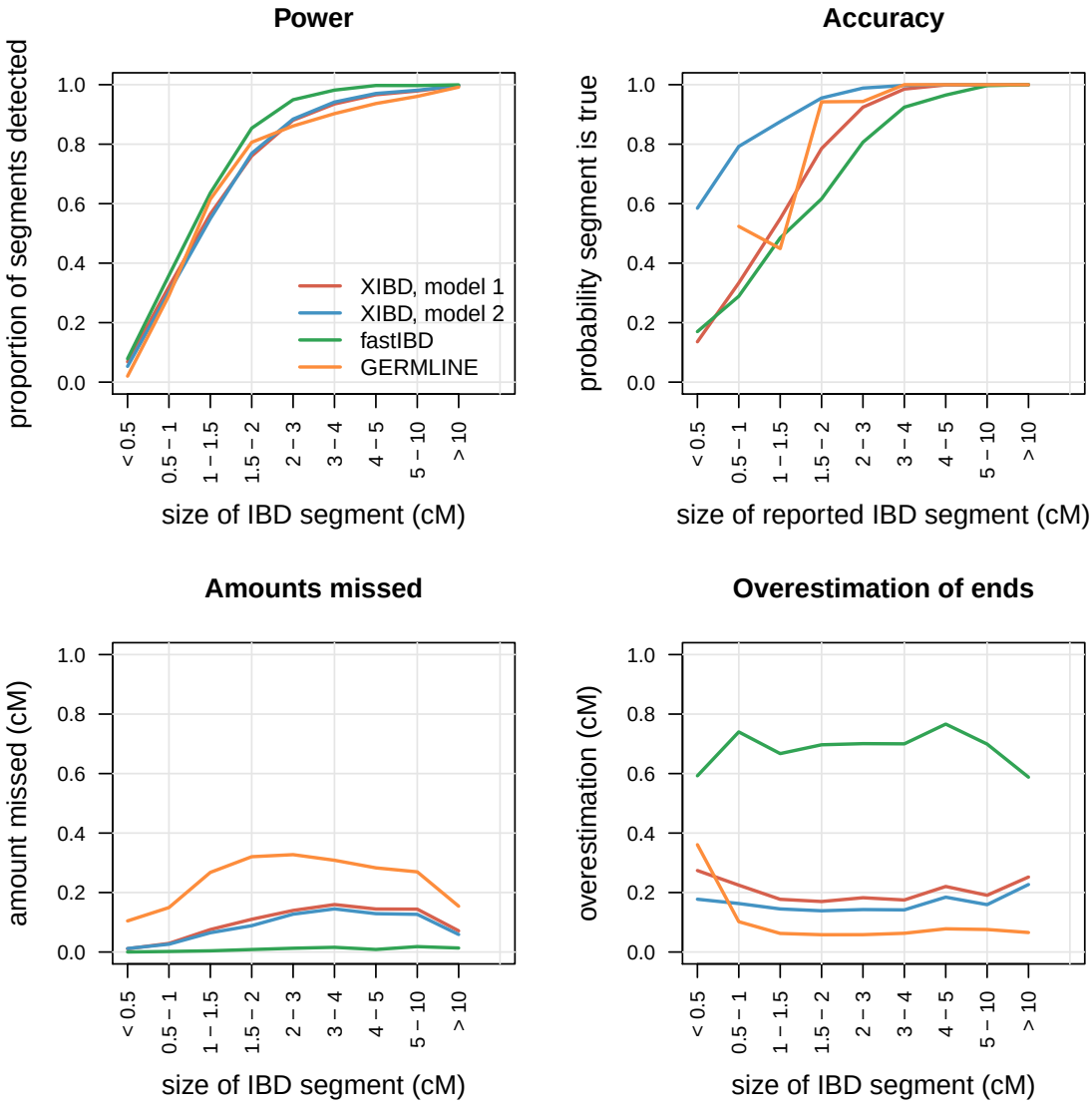


Figure 3.5: Comparison of IBD power, accuracy, under- and overestimation of XIBD, GERMLINE and fastIBD between 2 diploid chromosomes, calculated across various segment sizes.

3.3.2 Results

Performance results are displayed in Figures 3.5, 3.6 and 3.7, for pairs of diploid chromosomes, pairs of chromosomes where one chromosome is diploid and the other chromosome is haploid, and pairs of haploid chromosomes, respectively. When performing IBD analyses on diploid chromosomes, all algorithms have approximately the same level of power to infer segments of all sizes, however there are differences in the accuracy of each tool. XIBD model 2 (LD accounted for) is more likely to detect segments that are real than other tools. GERMLINE tends to underestimate segment lengths more than other tools, with on av-

erage 0.2 cM of a segment undetected, while fastIBD does not underestimate segments at all. This is likely due to fastIBD overestimating segments by considerable amounts (0.75 cM), resulting in little chance of underestimation of a segment's breakpoints, while GERMLINE only overestimates segment breakpoints by 0.1 cM on average.

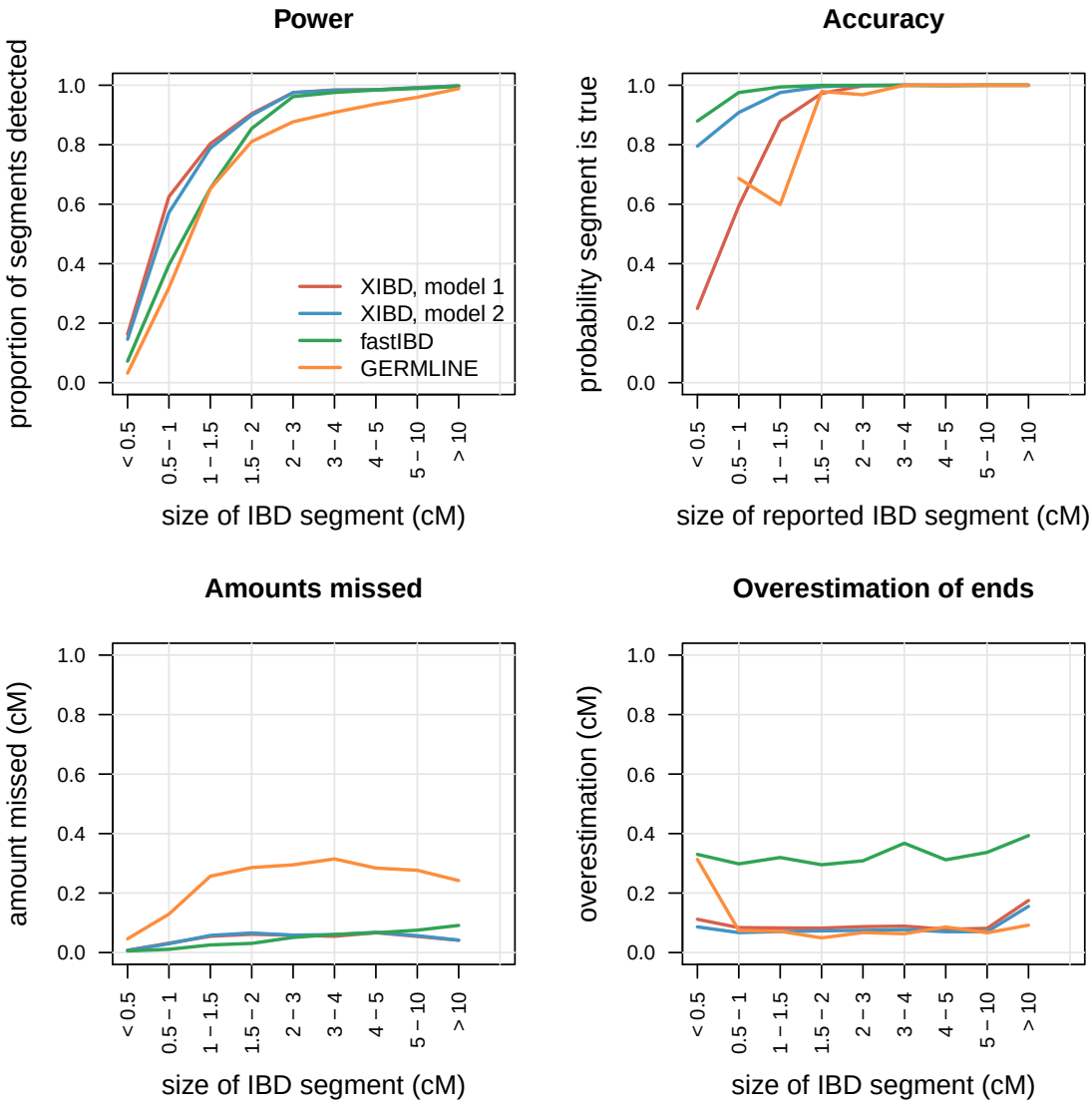


Figure 3.6: Comparison of IBD power, accuracy, under- and overestimation of XIBD, GERMLINE and fastIBD between between one diploid and one haploid chromosome, calculated across various segment sizes.

As haploid chromosomes are included in the analysis, the results from GERMLINE are very similar to analyses of diploid chromosomes, with the exception that accuracy increases for smaller segments as analyses include more haploid chromosome. XIBD achieves higher power and accuracy with segments estimated more accurately as haploid chromosomes

are included in the analysis, with analysis of entirely haploid chromosomes performing extremely well.

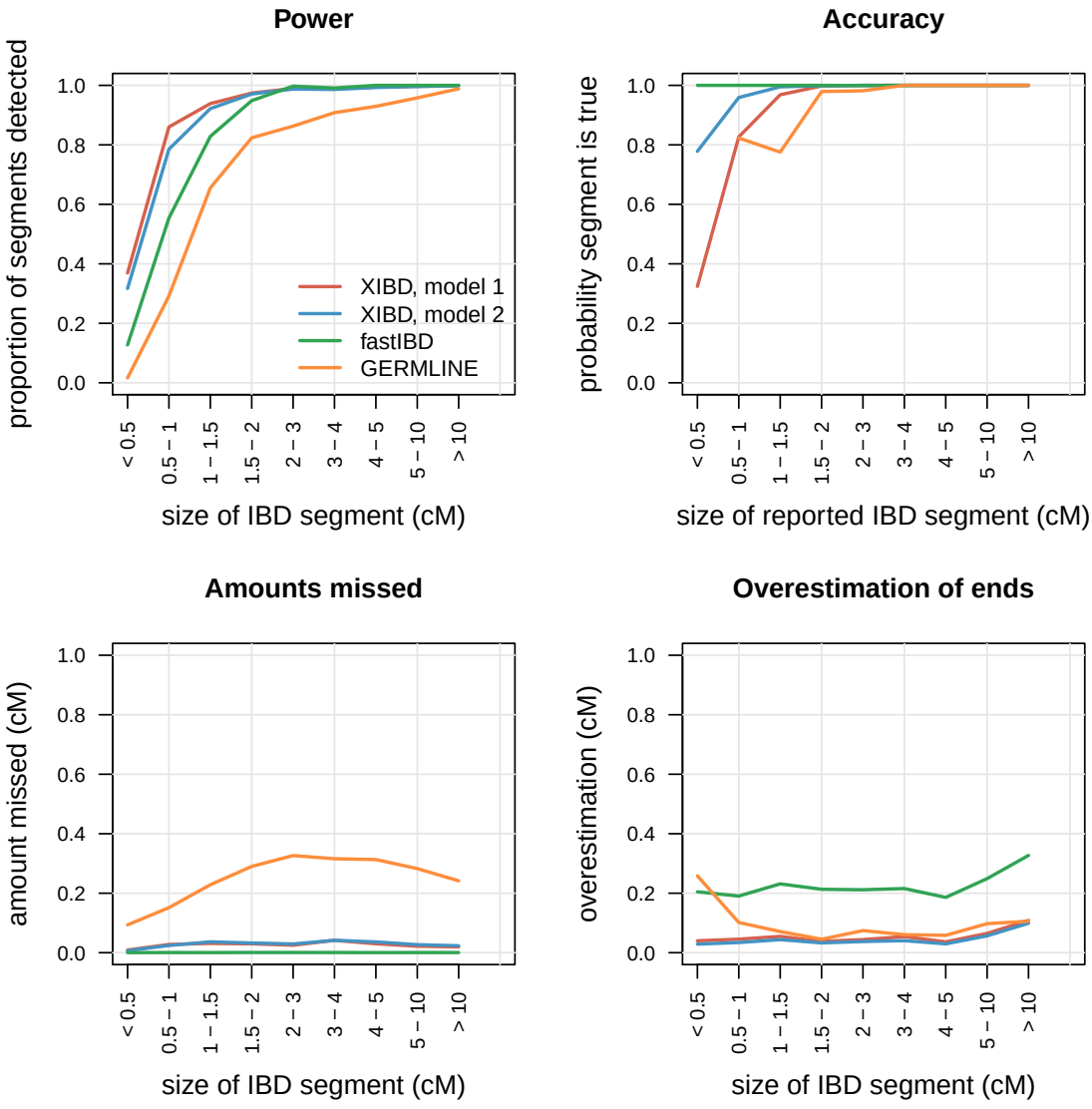


Figure 3.7: Comparison of IBD power, accuracy, under- and overestimation of XIBD, GERMLINE and fastIBD between 2 haploid chromosomes, calculated across various segment sizes.

The results from fastIBD are more dramatic as haploid chromosomes are included in the analysis. While the power of fastIBD increases slightly with the inclusion of haploid chromosomes, the accuracy of fastIBD increases markedly. FastIBD has the highest probability of detecting a segment that is real for all segments lengths. Furthermore, extremely small segments (< 0.5 cM) that are detected in cohorts of entirely haploid chromosomes are almost guaranteed to be real, although overestimation of segment breakpoints remains

high for fastIBD in analyses of all ploidy combinations.

3.3.3 Discussion

The overall performance of all tools increases with the inclusion of haploid chromosomes in analyses. This is most likely because there is no uncertainty surrounding haplotype phase for haploid chromosomes, which reduces noise in IBD estimation, allowing for more easily identified segments. Furthermore, reduced accuracy is expected for diploid chromosomes, since diploid chromosomes contribute two haplotypes to analyse, creating more opportunities for false detection of IBD segments.

The power of GERMLINE remains largely unchanged between haploid and diploid chromosomes, which is not surprising since phased haplotypes were used in the analysis of both diploid and haploid chromosomes. Therefore, the probability of detecting a segment should be approximately the same when analysing diploid or haploid chromosomes with known phase.

FastIBD performs exceptionally well on haploid cohorts, which is somewhat surprising given the BEAGLE HMM does not correctly account for haploid chromosomes. This implies that IBD analyses are largely robust to misspecification of probability distributions in terms of haploid chromosomes. While fastIBD outperforms all tools in terms of accuracy for extremely small segments in cohorts containing haploid chromosomes, it comes at a cost of overestimating segment boundaries.

XIBD performs consistently well across the four summary measures evaluated, and achieves the highest power for detecting small segments when haploid chromosomes are included in analyses. One important thing to note when evaluating these methods is that XIBD analyses were performed using unphased data, while GERMLINE used phased data and fastIBD phases data prior to IBD analysis. Thus, when considering diploid chromosomes, XIBD performance is comparable to methods that make use of haplotype phase information. If phased diploid data were available, one could run XIBD using a haploid model. This has the advantage of detecting homozygosity by descent (IBD between homologous chromosomes within an individual) with potentially improved power for IBD detection. However, we note that changes in power is exclusively determined by the accuracy of the phasing algorithm implemented.

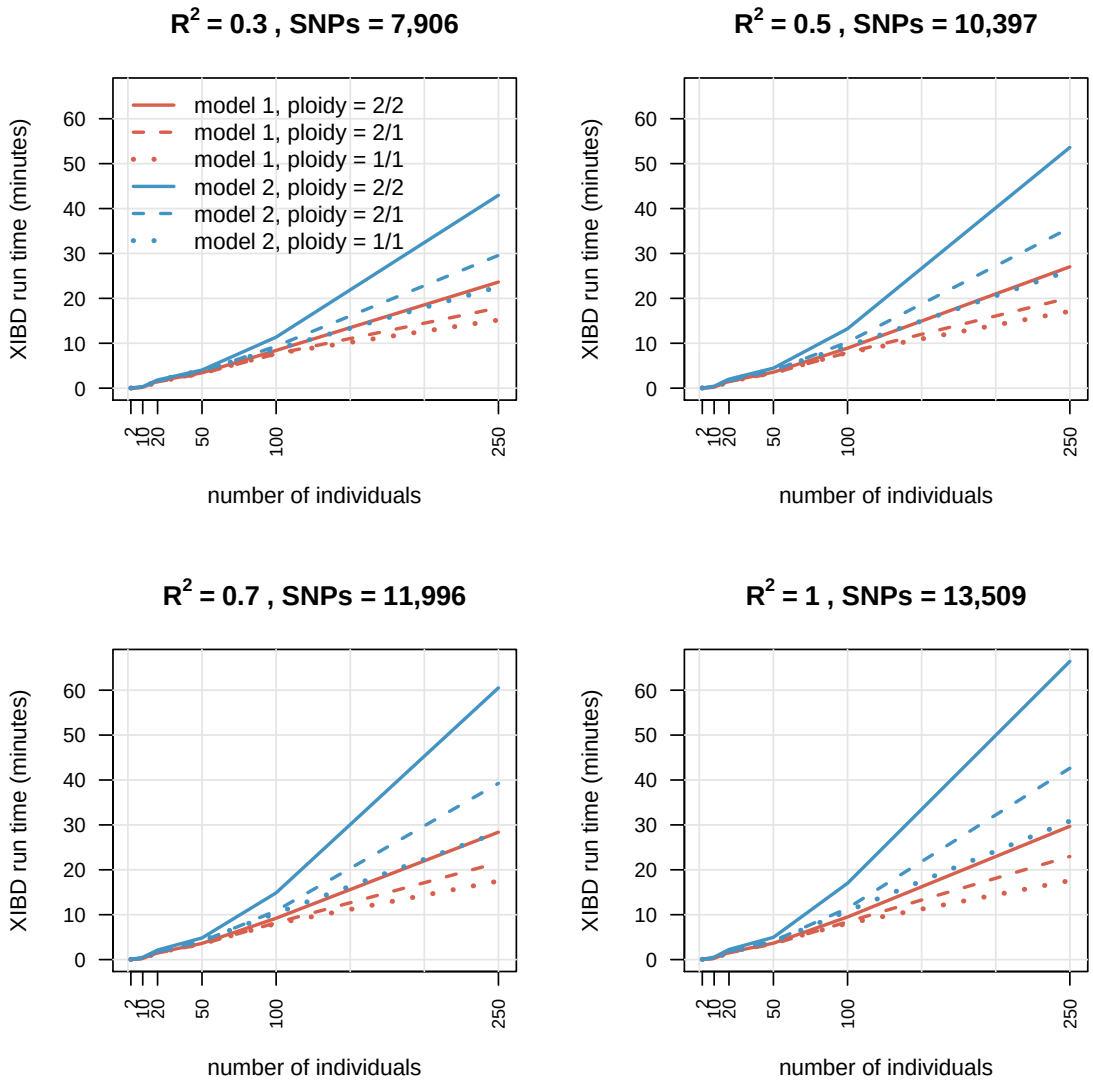


Figure 3.8: Computation times for XIBD on cohorts of varying sizes with SNPs in different states of LD. The cohort sizes of 2, 10, 20, 50, 100 and 250 individuals correspond to 1, 45, 190, 1125, 4950 and 31125 pairwise analyses respectively. Run times are the real elapsed run times as calculated by the proc.time function in R and the analysis was performed on a dual socket Xeon E5-2690v3, 512GB memory machine. XIBD was running using 5 cores.

Given that we are interested in recent common ancestry (up to 25 generations and 50 meiosis), the threshold on IBD segment lengths detected with high power and accuracy is 2 cM. However, our results show that, for XIBD and fastIBD at least, the threshold on IBD length can possibly be reduced for haploid chromosomes to 1 cM or 1.5 cM, identifying individual with more distant ancestry, potentially separated by up to 50 generations (100 meiosis).

3.4 Computation time of XIBD

HMMs are notorious for being computationally intensive algorithms to run, and XIBD is no exception. Computational time increases linearly with the number of SNPs and quadratically with the number of individuals (Figure 3.8), with an analysis of 250 individuals (31,125 pairwise analyses) and 13,509 SNPs taking slightly longer than 1 hour. Accounting for LD in model 2 increases the computational burden as conditional emission probabilities increase the complexity of the model. Furthermore, datasets which include diploid chromosomes take longer to analyse as the observation state space increases, resulting in more genotypic combinations to account for.

XIBD computation time can be reduced by running model 1 with potentially fewer SNPs at little cost to performance. Furthermore, XIBD allows parallelization of analyses on multicore processors. While XIBD is not computationally comparable to tools like GERMLINE and fastIBD, which took less than 5 seconds and less than 10 minutes to analyze 13,509 SNPs for 250 individuals, respectively, it is still a valuable tool for analysis of small to medium sized datasets containing tens to hundreds of individuals.

3.5 XIBD graphical representations of relatedness

The results from IBD analyses are typically returned as plain text files containing information on the genomic locations of inferred segments. These can be cumbersome to sift through if many IBD segments have been detected, particularly when cohorts contain mixed phenotypes. XIBD is implemented in R which allows us to make use of the graphical interface that R provides. As such, we have developed a number of graphical functions to facilitate with interpreting results. All figures are produced using ggplot2[76], unless otherwise stated.

3.5.1 Kinship confirmation using IBD coefficients

The first of such functions allows one to check that, for individuals with a known degree of relatedness, the reported relationship is likely to be correct. XIBD calculates identity coefficients Δ_k $(k = 1, \ldots, 9)$ for a given pedigree using IdCoefs[70] for autosomes and our implementation for the X chromosome (Appendix B), from which IBD coefficients, ω_0, ω_1 and ω_2, can be approximated. Following this, the theoretical proportion of genome shared IBD for each pair can be calculated as

$$\pi = \frac{\omega_1}{2} + \omega_2.$$

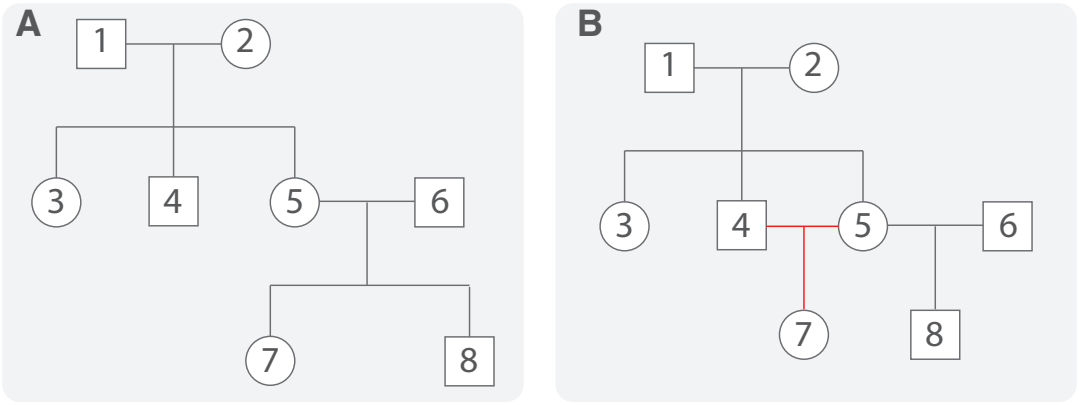


Figure 3.9: Two pedigrees for a 3 generation family containing 8 individuals. **A** An incorrectly reported pedigree for the family. **B** Correct pedigree showing consanguinity between individuals 4 and 5, where individual 7 is actually the daughter of individual 5, not individual 6. The red lines depict the incorrectly reported relationship.

SNP genotype data can be used to approximate π from IBD coefficients calculated in Chapter 2, without the use of pedigree information. The estimated values of π can be compared to the theoretical values in XIBD using a heat map to identify individuals with misspecified relationships. For example, consider the two pedigrees in Figure 3.9. Figure A represents the reported pedigree for a small family while Figure B represents the actual pedigree for the same family, which contains consanguinity between a pair of siblings. Identity coefficients were calculated for all pairs of individuals given the relationships specified in each pedigree from which values of π were calculated. These denote the theoretical IBD proportion for each pedigree. We also simulated values of π for all pairs assuming an approximate normal distribution with mean value of π_i for pair i in pedigree B (actual pedigree) and standard deviation between 0.01 and 0.1. These were taken to represent estimates from SNP genotype data.

Figure 3.10 displays a heat map of estimated and theoretical IBD proportions for each pedigree, for autosomes and the X chromosome respectively. The theoretical proportions should approximately match the estimated proportions, producing symmetrical figures, which is not the case for pedigree A. Specifically, the reported father (individual 6) of individuals' 7 and 8 is not predicted to be related to individual 7. In fact, individual 7 shares more of their genome than expected with all other individuals in this pedigree,

particularly individual 5. If we assume that individual 5 is the father of individual 7, as in pedigree B, we get approximately symmetrical heat maps, confirming the suspected relationships. Visual displays such as this allow for rapid assessment of moderate to large cohorts in terms of reported versus inferred pedigree relationships.

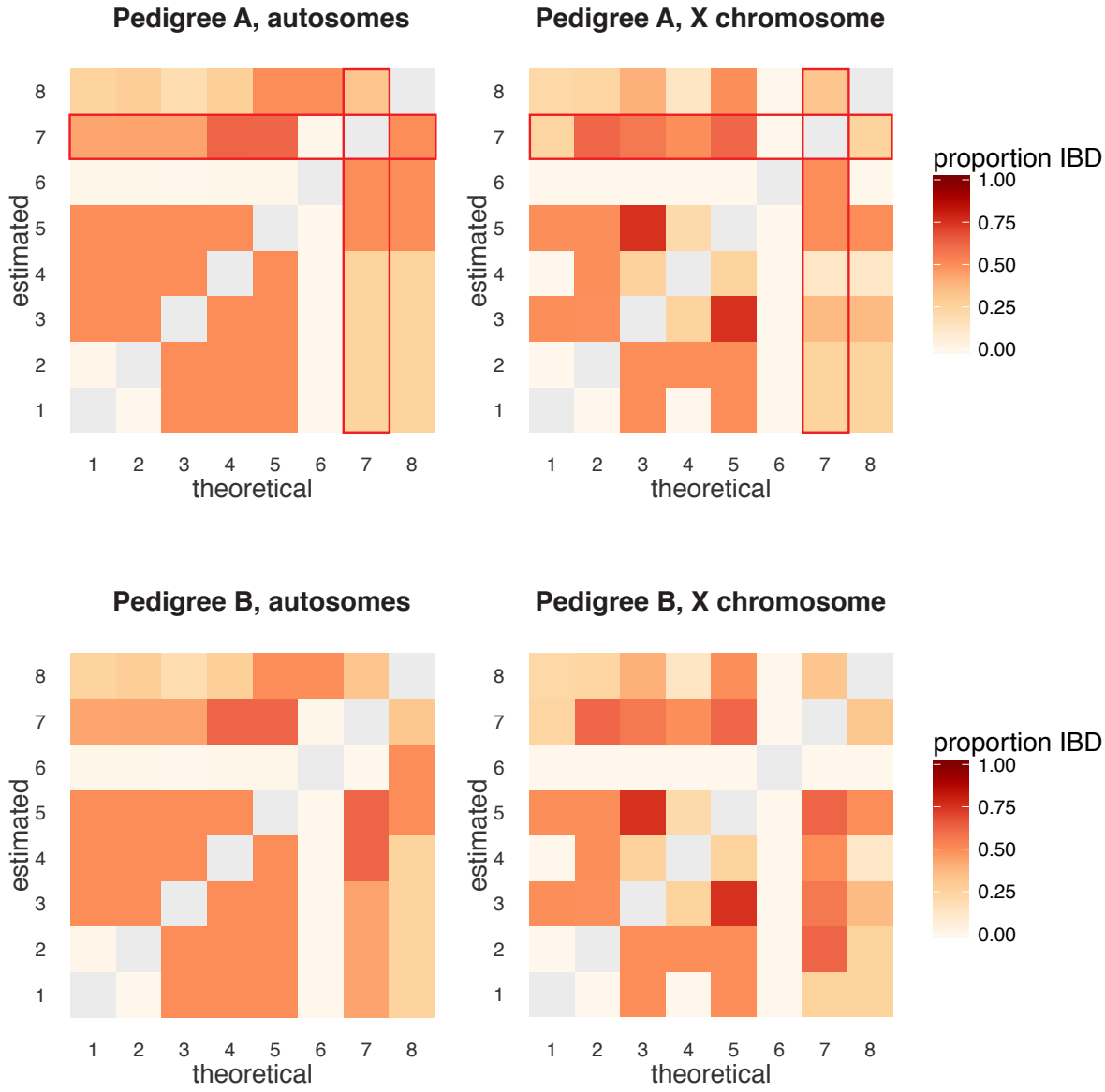


Figure 3.10: Example figures produced by XIBD displaying the approximate proportion of genome shared IBD between pairs of individuals, given a pedigree. Theoretical identity coefficients for autosomes were calculated using the IdCoef[70] R package while our implementation was used for the X chromosome. Estimated proportions are calculated from SNP genotype data as in Section X, although have been simulated for this example. Grey coloured boxes represent IBD proportions with oneself, which are not reported by XIBD and hence are NA values. All values of π estimated with individual 7 in pedigree A have been highlighted with red boxes as suspicious.

3.5.2 IBD segments and excess IBD

Following IBD analysis, inferred segments can be displayed in a brick-style plot to visualize the break points, IBD sizes and overlap of segments detected in multiple pairs (Figure 3.11). Gene annotations can be added to the figure which enables fast identification of individuals sharing a common haplotype over a gene(s) of interest.

When many pairs are inferred IBD, displaying segments as in Figure 3.11 may be impractical. Therefore, we provide an alternative to this function, which simply calculates the proportion of pairs who are IBD at each SNP and plots this distribution across the genome. Although this does not identify sharing between specific pairs, it is a quick way of identifying genomic loci with many IBD pairs. Genomic loci with many shared segments tend to be associated with natural selection.

In particular, Albrechtsen et al.[21] and Han and Abney[22] have used similar figures to successfully identify the HLA region as under positive selection in a number of populations, in addition to the human lactase gene (LCT). We calculated the proportion of pairs IBD at each SNP in the example data used in Figure 3.11 and display the results in Figure 3.12. A large proportion of pairs share segments of IBD on chromosome 20.

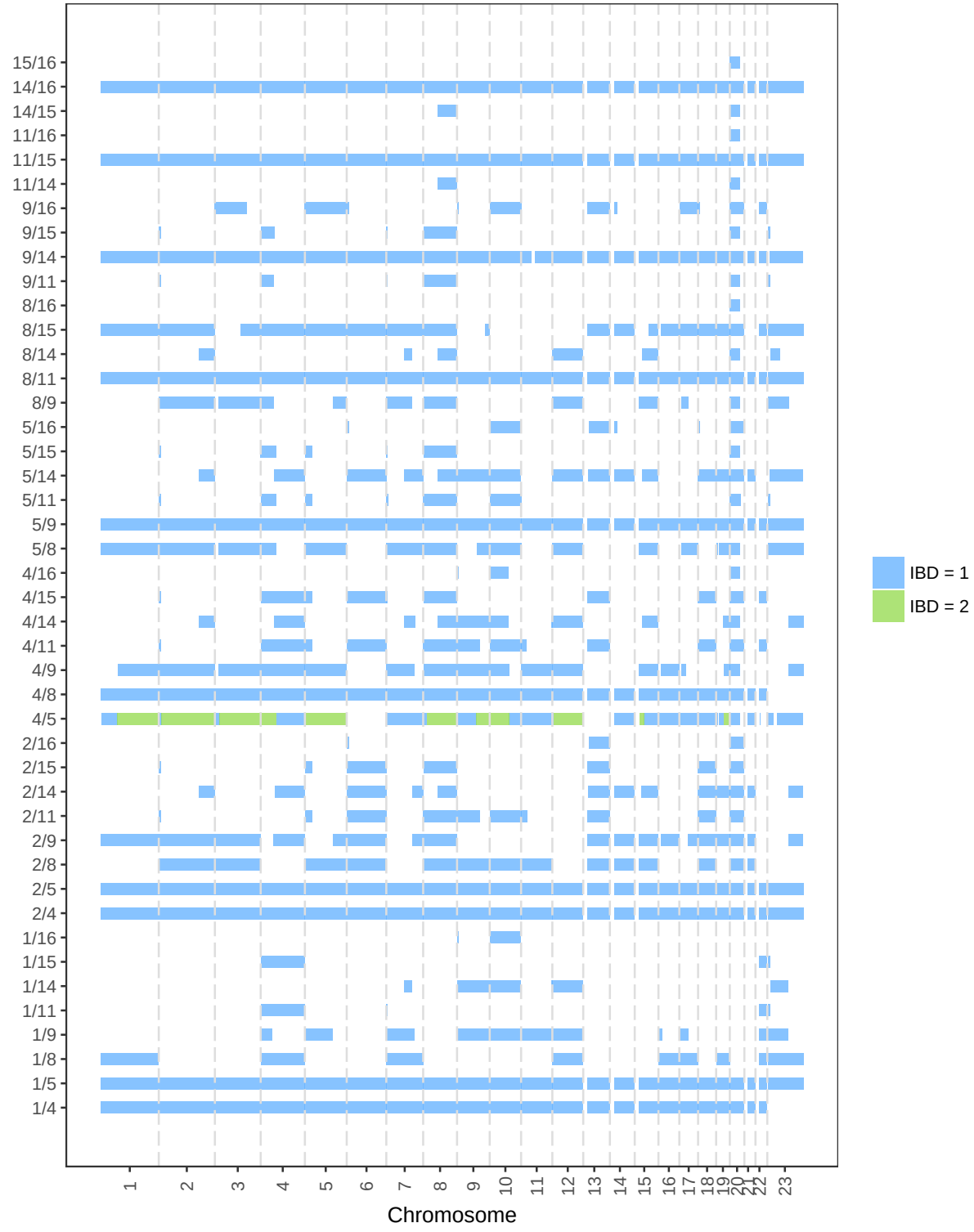


Figure 3.11: An example figure produced by XIBD displaying segments of IBD across the genome. Each coloured segment denotes an IBD region with either 1 or 2 alleles inferred IBD. Data used in this figure was simulated from a 5 generation pedigree according to algorithms 1 and 2, respectively, using SNPs extracted from the Illumina HumanOmni2.5 platform. A haplotype on chromosome 20 appears to have be inherited in most individuals.

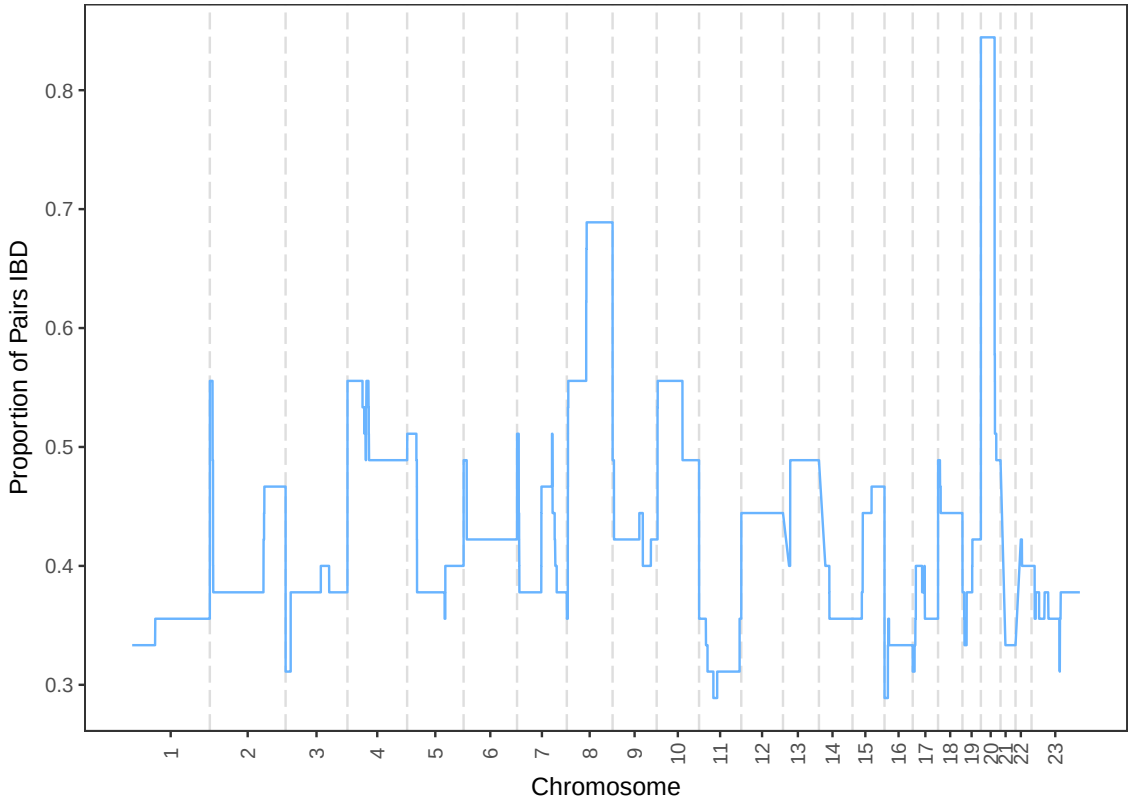


Figure 3.12: An example figure produced by XIBD displaying the proportion of pairs IBD at each SNP. The data used in this figure is the same as that used in Figure 3.11

3.6 Summary

Here we evaluate the performance of XIBD. We show that XIBD achieves high power and accuracy to infer IBD segments of 2cM and larger, and it also performs remarkably well for segments as small as 0.5cM when haploid chromosomes are included in the analysis. Furthermore, we demonstrate that XIBD is just as powerful, if not more powerful, than GERMLINE and fastIBD in many instances. Unfortunately, XIBD is computationally intensive, unlike GERMLINE and fastIBD, making it unsuitable for large cohorts with many hundreds or thousands of individuals. Nonetheless XIBD is a valuable tool for analysis of the X chromosome, and small cohorts in general, as it includes a number of graphical features that most IBD software do not, making for easier interpretation of results. We applied XIBD to several applications in human genetics, one of which is included as Chapter 4, while another is presented in Shaw et al. [19].

Chapter 4

XIBD application to an epilepsy cohort

4.1 Identity by descent fine mapping of familial adult myoclonus epilepsy (FAME) to 2p11.2-2q11.2

Hum Genet
DOI 10.1007/s00439-016-1700-8

ORIGINAL INVESTIGATION

Identity by descent fine mapping of familial adult myoclonus epilepsy (FAME) to 2p11.2–2q11.2

Lyndal Henden[1,2] · Saskia Freytag[1,2] · Zaid Afawi[3] · Sara Baldassari[4] · Samuel F. Berkovic[5] · Francesca Bisulli[6,7] · Laura Canafoglia[8] · Giorgio Casari[9] · Douglas Ewan Crompton[10] · Christel Depienne[11,12] · Jozef Gecz[13,14] · Renzo Guerrini[15,16] · Ingo Helbig[17,18,19] · Edouard Hirsch[20] · Boris Keren[21,22] · Karl Martin Klein[23,24] · Pierre Labauge[25] · Eric LeGuern[22,26,27] · Laura Licchetta[6,7] · Davide Mei[15] · Caroline Nava[21,22] · Tommaso Pippucci[4] · Gabrielle Rudolf[11,28] · Ingrid Eileen Scheffer[5,29,30] · Pasquale Striano[31] · Paolo Tinuper[6,7] · Federico Zara[32] · Mark Corbett[13] · Melanie Bahlo[1,2]

Received: 18 April 2016 / Accepted: 21 June 2016

Abstract Familial adult myoclonus epilepsy (FAME) is a rare autosomal dominant disorder characterized by adult onset, involuntary muscle jerks, cortical myoclonus and occasional seizures. FAME is genetically heterogeneous with more than 70 families reported worldwide and five potential disease loci. The efforts to identify potential causal variants have been unsuccessful in all but three families. To date, linkage analysis has been the main approach to find and narrow FAME critical regions. We propose an alternative method, pedigree free identity-by-descent (IBD) mapping, that infers regions of the genome between individuals that have been inherited from a common ancestor. IBD mapping provides an alternative to linkage analysis in the presence of allelic and locus heterogeneity by detecting clusters of individuals who share a common allele. Succeeding IBD mapping, gene prioritization based on gene co-expression analysis can be used to identify the most promising candidate genes. We performed an IBD analysis using high-density single nucleotide polymorphism (SNP) array data followed by gene prioritization on a FAME cohort of ten European families and one Australian/New Zealander family; eight of which had known disease loci. By identifying IBD regions common to multiple

Electronic supplementary material The online version of this article (doi:10.1007/s00439-016-1700-8) contains supplementary material, which is available to authorized users.

✉ Melanie Bahlo
bahlo@wehi.edu.au

1 Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC 3052, Australia

2 Department of Medical Biology, University of Melbourne, Melbourne, VIC 3010, Australia

3 Tel Aviv University Medical School, 69978 Tel Aviv, Israel

4 Medical Genetics Unit, Polyclinic Sant'Orsola-Malpighi-Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy

5 Epilepsy Research Centre, Department of Medicine, University of Melbourne Austin Health, Melbourne, VIC 3084, Australia

6 IRCCS Istituto delle Scienze Neurologiche di Bologna, Bologna, Italy

7 Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy

8 Neurophysiopathology and Epilepsy Center, IRCCS Foundation C. Besta Neurological Institute, Milan, Italy

9 Division of Genetics and Cell Biology, Università Vita-Salute San Raffaele, San Raffaele Scientific Institute, Milan, Italy

10 Neurology Department, Northern Health, Melbourne, VIC 3076, Australia

11 Département de Médicine translationnelle et Neurogénétique, IGBMC, CNRS UMR 7104/INSERM U964/Université de Strasbourg, Illkirch, France

12 Laboratoire de diagnostic génétique, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

13 Robinson Institute and School of Medicine, The University of Adelaide, Adelaide, SA 5005, Australia

14 School of Biological Sciences, The University of Adelaide, Adelaide, SA 5005, Australia

15 Pediatric Neurology, Neurogenetics and Neurobiology Unit and Laboratories, Neuroscience Department, A Meyer Children's Hospital, University of Florence, Florence, Italy

16 IRCCS Stella Maris Foundation, Pisa, Italy

17 Department of Neuropediatrics, Christian-Albrechts-University of Kiel and University Medical Center, Kiel, Schleswig-Holstein, Germany

Published online: 01 July 2016

families, we were able to narrow the FAME2 locus to a 9.78 megabase interval within 2p11.2–q11.2. We provide additional evidence of a founder effect in four Italian families and allelic heterogeneity with at least four distinct founders responsible for FAME at the FAME2 locus. In addition, we suggest candidate disease genes using gene prioritization based on gene co-expression analysis.

Introduction

Familial adult myoclonus epilepsy (FAME), also known as familial cortical myoclonic tremor with epilepsy (FCMTE), autosomal dominant cortical myoclonus and epilepsy (ADCME), and benign adult familial myoclonic epilepsy (BAFME), is a rare autosomal dominant disorder characterized by rapid involuntary muscle jerks of cortical origin mimicking essential tremor and sporadic seizures affecting more than 70 families worldwide (Licchetta et al. 2013). Onset of the disorder can range between 10 and 60 years of age and is typically slowly progressing or non-progressive and non-disabling (Crompton et al. 2012).

FAME was first described in two individuals of Japanese descent (Ikeda et al. 1990). Thereafter, a number of Japanese families were reported with features of the disorder and genomic analyses identified linkage to a 7.16 megabase (Mb) region spanning chromosome 8q24 (FAME1) (Mikami et al. 1999; Plaster et al. 1999; Mori et al. 2011). Linkage to this region has only been identified in families of Japanese descent and as yet no causal variants have been identified.

Elia et al. (1998) and Guerrini et al. (2001) were first to report European families with characteristics of FAME in addition to mild-to-moderate intellectual disability. Guerrini et al. (2001) performed a genomic analysis and identified a second disease locus (FAME2) on chromosome 2. Following this, multiple European families with FAME characteristics have been described and the FAME2 locus was refined to a 10.4 Mb region spanning chromosome 2p11.1–q12.2 and containing 61 RefSeq genes (Madia et al. 2008; Saint-Martin et al. 2008; Crompton et al. 2012; Licchetta et al. 2013). de Fusco et al. (2014) sequenced candidate genes within the FAME2 locus and identified a novel in-frame insertion/deletion in the *ADRA2B* gene potentially causal for the disorder in two families from Tuscany, Italy. This gene was also sequenced in a number of families from southern Italy linked to the FAME2 locus; however, mutations in the *ADRA2B* gene were not detected. The FAME2 locus overlaps the centromere of chromosome 2. Centromeres undergo suppressed recombination and usually contain large gaps in marker information (Choo 1998). This makes the FAME2 critical region especially difficult to assess; likely contributing to the lack of success in finding causal variants.

A third disease locus (FAME3) was mapped to chromosome 5p15.31–5p15.1 in a single large French family (Depienne et al. 2010), and more recently, a fourth disease locus was reported on chromosome 3q26.32–3q28 (FAME4) in a Thai family (Yeetong et al. 2013). In addition to this, a consanguineous Egyptian family with features of FAME has been reported to have a homozygous mutation in the *CNTN2* gene on chromosome 1q32 (FAME5) (Stogmann et al. 2013).

A founder effect has been proposed for six families, residing within close proximity in Italy, that have been mapped to the FAME2 locus and have a likely, identical microsatellite marker-based haplotype segregating in all

[18] Departments of Brain and Cognitive Sciences, Physiology and Cell Biology, Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Negev, Israel

[19] Division of Neurology, The Children's Hospital of Philadelphia, Philadelphia, USA

[20] Medical and Surgical Epilepsy Unit, Hautepierre Hospital, University of Strasbourg, Strasbourg, France

[21] Département de Génétique, Hôpital de la Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, 75013 Paris, France

[22] Sorbonne Universités, UPMC Univ Paris 06,UMR S 1127, ICM, 75013 Paris, France

[23] Department of Neurology, Epilepsy Center Frankfurt Rhine-Main, Center of Neurology and Neurosurgery, University Hospital, Goethe-University Frankfurt, Frankfurt, Germany

[24] Department of Neurology, Epilepsy Center Hessen, University Hospitals Giessen and Marburg, Philipps-University Marburg, Marburg, Germany

[25] Department of Neurology, Montpellier University, Gui de Chauliac, 34295 Montpellier, Cedex 5, France

[26] INSERM, U 1127; CNRS, UMR 7225; INSERM UMR 975; Institut du Cerveau et de la Moelle Epinière; and Département de Génétique et de Cytogénétique, Hôpital de la Pitié-Salpêtrière, Assistance Publique–Hôpitaux De Paris (AP-HP), Paris, France

[27] Université Pierre et Marie Curie (Paris 6) (UPMC), UMRS 975, Paris, France

[28] Department of Neurology, Hautepierre Hospital, University of Strasbourg, Strasbourg, France

[29] Florey Institute of Neuroscience and Mental Health, Melbourne, VIC 3084, Australia

[30] Department of Paediatrics, University of Melbourne, Royal Children's Hospital, Melbourne, VIC 3052, Australia

[31] Pediatric Neurology and Muscular Diseases Unit, Department of Neurosciences, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health, Gaslini Institute, Genoa, Italy

[32] Laboratory of Neurogenetics, Department of Neurosciences, Gaslini Institute, Genoa, Italy

Table 1 Description of the FAME cohort data set

Family	Publication	Origin	Number of Affected individuals	Disease locus	Possible pathogenic variant
Family 1	Guerrini et al. (2001), De Fusco et al. (2014)	Italian province of Livorno, Tuscany	11	2p11.1–q12.2	*ADRA2B*
Family 2	Striano et al. (2005)	Italian province of Caserta, Campania	18	2p11.1–q12.2	–
Family 4	Licchetta et al. (2013)	Italian province of Caserta, Campania	25	2p11.1–q12.2	–
Families 3, 5	–	Italian province of Caserta, Campania	12, 4	2p11.1–q12.2	–
Family 6	Saint-Martin et al. (2008)	Spain	13	2p11.1–q12.2	–
Family 7	Crompton et al. (2012)	Australia/New Zealand	55	2p11.1–q12.2	–
Family 8	Depienne et al. (2010)	France	16	5p15.31–p15	–
Families 9, 10, 11	–	France	3, 7, 2	–	–

families (Madia et al. 2008; Licchetta et al. 2013). However, the identification of a pathogenic variant has been unsuccessful so far. In addition to this, de Fusco et al. (2014) proposed a founder effect in two more families from Italy with an identical variant in *ADRA2B*, within FAME2, based on shared haplotypes. This variant was not present in other families linked to FAME2.

To date, linkage analyses have been the primary method of analysis applied to FAME data sets to narrow the search space for causal variants. We propose an alternative method, identity-by-descent (IBD) mapping, which infers that the regions of the genome have been inherited from a common ancestor and can be applied to cohorts of unrelated individuals without a pedigree (Browning and Browning 2010). IBD analyses can lead to the identification and reduction in size of the critical regions in which likely causal variants should be located. In addition to this, they can lead to the discovery of distant relatedness between families unknown to be related (Albrechtsen et al. 2009).

Whole exome sequencing (WES) and whole genome sequencing (WGS) have failed to identify a variant for FAME thus far. This could be attributed to a variety of reasons, including the use of a poor reference genome with missing genetic information (Anvar et al. 2014); disease-causing variants are non-coding SNPs or splice-site mutations (Koboldt et al. 2013). To identify candidate disease genes within FAME loci, we propose in silico gene prioritization. In silico prioritization aims at discovering candidate disease genes by making use of deposited gene expression data and known disease-causing genes for the same or a related disorder to build a network that can be used to rank candidate genes (Aerts et al. 2006; Oliver et al. 2014). To date, no reported gene prioritization has been performed on candidate genes within the identified FAME critical regions.

Materials and methods

Patients and families

Eleven families were recruited for the analysis according to the relevant ethics of each country. Of these, 6 families have their clinical features described elsewhere (see Table 1). Families 1 to 7 have had their disease locus mapped to 2p11.1–q12.2, family 8 has been mapped to 5p15.31–p15.1, while families 9, 10 and 11 have unknown disease loci. Additionally, family 1 has a suspected disease variant in the *ADRA2B* gene while family 2 has had variants in *ADRA2B* excluded as causal (de Fusco et al. 2014). Two affected individuals were selected from each family as representatives for further analyses to determine relatedness and thus the extent of the locus and allelic heterogeneity in this cohort of FAME families. There were no known relationships between individuals from different families, although relatedness had been hypothesized between families 2, 3, 4, and 5 using microsatellite marker allele-based haplotype sharing comparisons.

All 22 individuals were SNP genotyped at the Département de Génétique Hôpital Pitié-Salpétrière using the Illumina HumanCytoSNP-12 chip, as already described (Nava et al. 2014). This chip has over 290,000 SNPs with a median physical map distance of 6.2 Kb (first and third quartiles of 4.3 and 14.3 Kb, respectively) and a median map distance of 0.004 cM (first and third quartiles of 0.001 and 0.013 cM, respectively). SNP genotype calls were generated using the Illumina's GenomeStudio Software.

Data processing and IBD methodology

HapMap Tuscan (TSI) allele frequency data were used in the analysis, after this HapMap population was found to be

a good fit (The International HapMap 3 Consortium 2010) (Supplementary Material). We used XIBD (Henden et al. 2016) to estimate relatedness and infer genomic regions shared IBD between pairs of individuals. Briefly, XIBD implements a first-order continuous-time hidden Markov model to detect IBD using unphased genotype data and can account for linkage disequilibrium (LD), genotyping errors, and missing data. The Viterbi algorithm is used to find the most likely sequence of IBD states between a pair of individuals, where the states are the number of alleles-shared IBD with possible values of 0, 1, or 2 alleles (Rabiner 1989).

XIBD model 2 was run which uses haplotype frequencies to account for LD. The HapMap Phase 3 TSI genotype data were used as a reference data set to calculate the haplotype frequencies and allele frequencies. Markers in near perfect LD (R^2 >0.99) were removed from the analysis along with markers that had low minor allele frequencies (MAF <0.01) and markers with missing data for more than 50 % of individuals; this included seven adjacent SNPs overlapping the FAME2 critical region with missing genotype calls for all 22 individuals. Finally, IBD segments less than 0.5 cM or containing less than 20 SNPs were not considered as candidates, as they represent population-background LD or deeper ancestral LD which will lead to IBD inference that is not of relevance to the disease. This has been previously noted and discussed (Brown et al. 2012).

Gene prioritization using a combined resource of human brain expression data

The microarray gene expression data sets used for the in silico gene prioritization were all generated with tissues from population-based samples of post-mortem human brains. Supplementary Table 1 includes a detailed description of the five publicly available data sets that were used in combination to assess the network proximity for the candidate genes (in silico prioritization). Since the data sets were very different in their sampling design and technology used, advanced data cleaning was applied using RUVcorr (Freytag et al. 2015) (Supplementary Material). We used 421 known epilepsy genes (listed in Supplementary Table 2) to form the basis for defining putative relevant epilepsy genetic co-expression networks against which we assess our list of candidate FAME genes. Each of the known epilepsy genes was available in at least one data set. Finally, we removed all expression data derived from the prenatal brains, as it has been previously shown that brain gene expression patterns are very different from post-natal and adult brains, with greater variability (Kang et al. 2011). We argued that since FAME is mid-to-late age onset, genes of relevance would be best identified by focusing on post-natal brain expression which is dominated in the six data sets by adult brains (90 % of arrays that are >20 years old).

We chose to perform gene prioritization on candidate disease genes in the FAME1, FAME2, FAME3, and FAME4 critical regions, hypothesizing that the disease genes from the four loci should all be affecting the same pathway and hence be in the same network.

To prioritize the candidate disease genes, we constructed a network for every combination of candidates (one for each of the FAME1, FAME2, FAME3, and FAME4 loci) with the known epilepsy genes. Network construction was based on a weighted Pearson correlation coefficient combined with a thresholding approach similar to the approach in Oliver et al. (2014). In particular, we weighted each sample by the inverse of the squared number of samples that were also extracted from this brain. This approach takes into account that some brains have several hundred samples, while some brains only have one sample. Only correlations exceeding 0.60 were deemed to represent true co-expression. The thresholding approach allowed the construction of an adjacency matrix (a binary matrix indicating the presence or the absence of co-expression) corresponding to the network for every combination of genes. Note that due to the unavailability of some gene combinations, we had to assume that the genes in question did not interact, as otherwise, we would have had to considerably reduce the number of investigated genes.

Standard in silico gene prioritization approaches consider only single gene candidates. Here, we have combinations of four causal genes that we wish to identify, one from each of the four loci. This leads to a much longer candidate list. Therefore, using the adjacency matrices, we performed filtering to determine the most likely candidate genes and then prioritized these. We first filtered by removing combinations of genes that were not directly, or indirectly, interacting, i.e., they were not a member of the same pathway. Following this, we removed all gene combinations where the four candidate genes did not belong to the same cluster and did not share at least one neighbor. These conditions assume that all four candidate genes regulate the same gene in the pathway connected to the disease. Thus, we assume that dysfunctional regulation of the gene shared as a neighbor by all candidate genes leads to the pathway breaking down. This filtering procedure makes strong assumptions regarding the function of these genes, but this is necessary to reduce the number of candidate genes. We ordered the results of the final filtrations according to the sum of the PageRank [an algorithm for determining node importance (Brin and Page 1998)] of the gene combination in question.

Results and discussion

Assessing relationships

Pairwise IBD analyses were performed for all individuals in the data set. The individuals from families 1, 3, 9 and 11 are parent-offspring pairs. With the exception of consanguinity, a parent-offspring pair is an uninformative relationship to assess for IBD as one allele will be shared IBD across all autosomes; providing no opportunity to identify a critical region between these pairs alone. However, parent-offspring pairs can be useful for validating IBD between families, since they act as replicates, and thus add value to the analysis. More distant relatives would have been desirable for this analysis; however, sample availability and consent limited pair selection. Pairs chosen from the remaining seven families are cousins of the first degree or higher. We also inferred some families to be distantly related to the closest pair estimated as second cousins between families 8 and 10 (see Supplementary Table 3 for a full list of estimated parameters). The number of meiosis and initial probabilities of IBD sharing for the estimated relationships were used in the model to detect IBD tracts rather than those calculated from kinship coefficients for the pedigree relationships, as these were more likely to be accurate.

Inferred IBD tracts

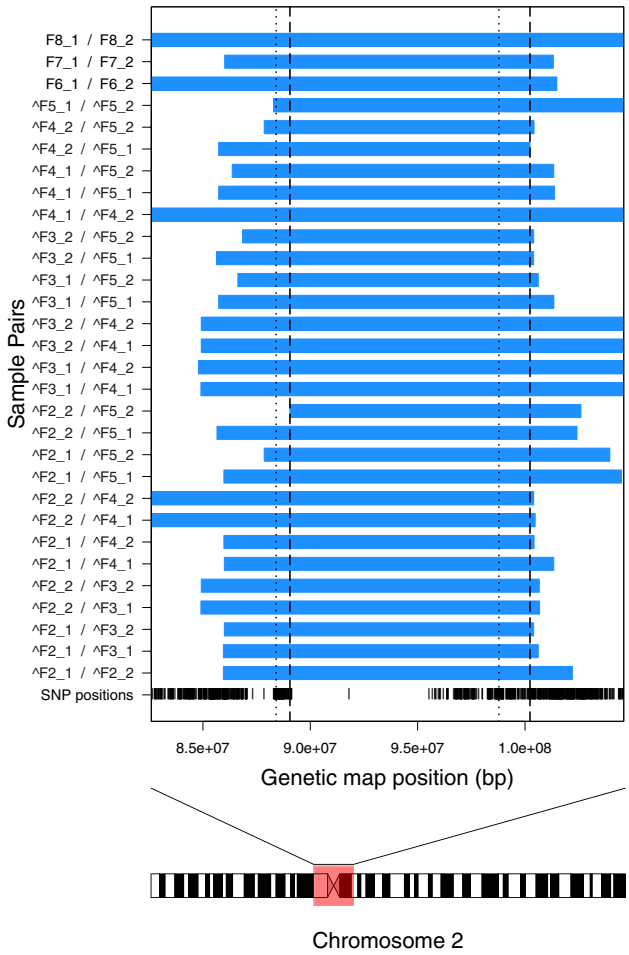


Fig. 1 IBD results across part of chromosome 2, drawn using R. The *y-axis* shows that the pair identifiers, e.g., *F2_1*/*F3_1*, represent the inferred IBD tracts between family *2*, individual *1* and family *3*, and individual *1*. The *x-axis* displays the genetic map position in base pairs along chromosome 2 using hg19 reference. *Solid blue horizontal rectangles* are regions, where pairs share one allele IBD, while *checkered yellow rectangles* represent two alleles-shared IBD (none detected). The *dotted vertical lines* mark the FAME2 linkage region as in Licchetta et al. (2013). The *dashed vertical lines* mark an IBD region common to all pairs with IBD inferred over the FAME2 locus. The *wedged labels* on the *y*-axis highlight all pairwise results from four Italian families (families *2–5*) suspected of having a common founder. Positions of SNPs included in the analysis can be seen below the IBD tracts along with an ideogram highlighting the interval plotted. For ease of interpretation, all parent-offspring IBD segments have been excluded from the figure; specifically, IBD within the pairs from families 1, 3, 9, and 11 (colour figure online)

2p11.2–q11.2

Our analysis identified several IBD tracts shared within families as well as between families over the FAME2 critical region (Figs. 1, 2; Supplementary Table 4). Families 1–7, which had been previously mapped to this region, had IBD tracts inferred that span part of or the entire FAME2 interval. Families 9 and 11 also have IBD tracts overlapping FAME2; however, it is unclear whether this is the critical region for these families due to the uninformative nature of their relationships. We could, however, exclude FAME2 as the critical region for family 10, as no IBD tract was inferred over this region.

All pairwise IBD analyses between families 2, 3, 4, and 5 resulted in shared IBD tracts over the FAME2 critical region (Fig. 1, wedged *y*-axis labels), suggesting relatedness of these four Italian families through a common ancestor. All four families are either from, or reside in close proximity to, Naples, Italy, and this result is consistent with the previous findings of a shared haplotype between Neapolitan families by identity by the state analysis using microsatellite data ($N = 4$ markers) (Madia et al. 2008; Licchetta et al. 2013). Relatedness between these families suggests that the same causal variant should be responsible for the FAME disorder in these cases.

There was no evidence of IBD sharing between families 1, 6, and 7 with any other family previously mapped to the FAME2 locus. This suggests that there are likely to be at least four distinct founders at the FAME2 locus

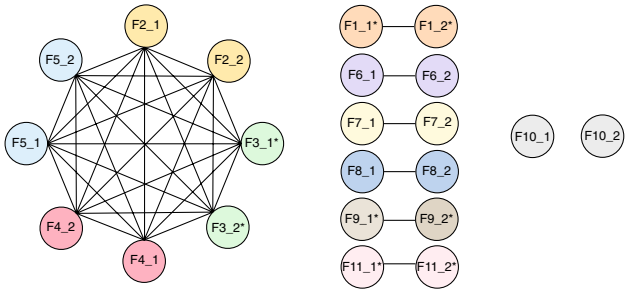


Fig. 2 IBD network of relatedness between all individuals over the FAME2 critical region, drawn using Adobe Illustrator CS6. Each node identifies one individual and *node colours* are unique for each family. An edge is drawn between two nodes if an IBD segment was inferred between the two individuals. Parent-offspring pairs are denoted by an *asterisk* within their nodes (families *1*, *3*, *9*, and *11*) (colour figure online)

corresponding to family 1 (Tuscany), the Neapolitan families 2, 3, 4, and 5, family 6 (Spanish), and family 7 (Australasian). Figure 2 displays an IBD network of relatedness between all individuals over the FAME2 critical region.

We investigated IBD tracts overlapping in multiple individuals within the FAME2 critical region. Taking the intersection of all IBD tracts produces an 11.2 Mb region-shared IBD between markers rs10179529 and rs1357719 (Fig. 1, dashed vertical lines). Towards, the start of this interval is a 6.56 Mb region containing only a single marker out of 203 IBD markers in the interval. This 6.56 Mb region coincides with the centromere of chromosome 2 and affects our ability to potentially refine the IBD boundary further.

Guerrini et al. (2001) first defined the FAME2 locus as 12.4 cM (21.3 Mb) spanning markers D2S2161 to D2S1897 within 2p11.1–q12.2 using an unspecified human genome build. Subsequent to this, using hg19 coordinates, the FAME2 critical region was refined several times to the most recent interval of 10.4 Mb spanning markers D2S2216 and D2S2175 within 2p11.2–q11.2 (Fig. 1, dotted vertical lines) (Madia et al. 2008; Saint-Martin et al. 2008; Crompton et al. 2012; Licchetta et al. 2013). Combining our IBD region with the most resent FAME2 interval further shortens the FAME2 critical region to 9.78 Mb region spanning markers rs10179529 and D2S2175 within 2p11.2–q11.2 and containing 53 RefSeq genes (https://genome.ucsc.edu). Table 2 provides a summary of the new FAME2 critical region and Fig. 3 shows the refinement of FAME2 over time and its current physical map location. We note that the *ADRA2B* gene remains in the critical region.

5p15.31–p15.1

We identified 3 IBD tracts overlapping the FAME3 critical region, in addition to the four parent-offspring IBD segments (Fig. 4, Supplementary Table 5). Of the families with IBD tracts inferred here, only family 8 has been linkage mapped to this region. This pair of individuals is inferred second cousins, and although they share only 10 % of their genome IBD, we were unable to narrow the FAME3 critical region further, as their IBD tract overlaps the critical region entirely (Depienne et al. 2010).

Our analysis also identified an IBD tract between the pair from family 4 and another tract between the pair from family 10. We believe that family 4 shares the same causal variant as families 2, 3, and 5 who have not been inferred IBD here; hence, we exclude FAME3 as a critical region for family 4 and cannot use this IBD segment to narrow the FAME3 critical region. Alternatively, family 10 has an unknown disease locus, so we cannot exclude FAME3 as a critical region for this family. The two individuals from family 10 are first cousins and share approximately 20 % of their genome IBD. We expect the causal variant for this family to be located in an IBD region; therefore, the search space has been greatly reduced.

Gene prioritization

There were 42, 53, 58, and 149 RefSeq genes identified within the FAME1, updated FAME2, FAME3, and FAME4 critical regions, respectively (see Supplementary Table 6). Using the combined microarray data sets, where gene names were standardized, we could identify 32 of the 42 genes within FAME1; 41 of the 53 genes within FAME2; 34 of the 58 genes within FAME3, and 91 out of the 149 genes within FAME5 (see Supplementary Table 6). After the first filtration step, we reduced the number of combinations of genes (one from each of the FAME1, FAME2, FAME3, and FAME4 loci) from over 4 million to 161,302. Assuming that all four candidate genes are present in the same co-expression network and regulate a common gene resulted in 61 prioritized gene combinations (see Supplementary Table 7).

The gene prioritization list contains a small number of genes that appear in many combinations. In particular, 9, 8, 4, and 13 genes were prioritized from the FAME1, FAME2, FAME3, and FAME4 loci, respectively.

Table 2 Description of IBD refined FAME2 critical region

Start marker	End marker	Start position (bp)	End position (bp)	Length (Mb)	No. of markers	No. of RefSeq genes
rs10179529	D2S2175	89,001,656	98,784,776	9.78	126	53

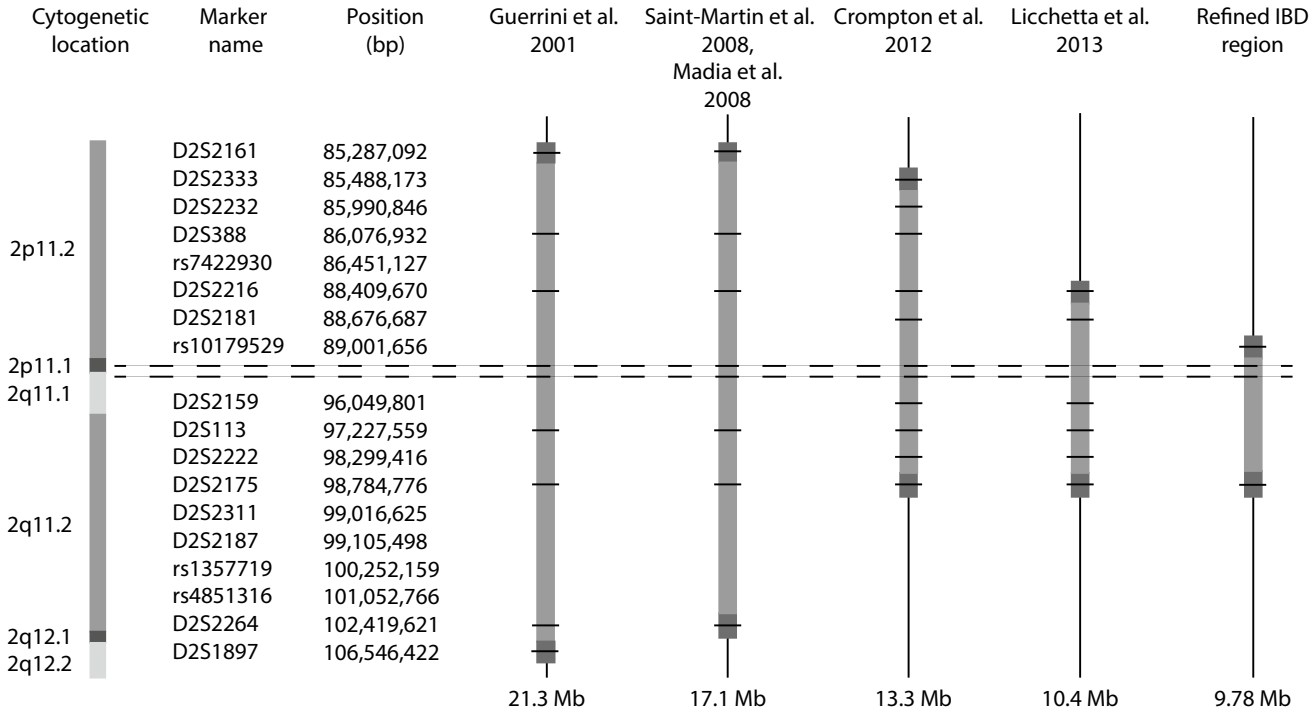


Fig. 3 Refinement of the FAME2 critical region over time, drawn using Adobe Illustrator CS6. The *darker* segments mark the boundaries of the reported intervals, and *solid horizontal lines* through the critical regions represent the microsatellite markers in the interval, which were included in the analysis. SNPs are not indicated in the IBD region derived from the SNP array data due to the large number of markers (126 SNPs). The *double black dashed horizontal lines* spanning the entire figure indicate the centromere. The base pair positions for makers starting with 'D2' are from http://rgd.mcw.edu (hg19), and the positions of the markers beginning with 'rs' are from HapMap Phase 3 data using hg19

Extensively sequenced and analyzed genes *ADRA2B* and *KCNIP3*, located within the FAME2 locus, appear in multiple gene combinations. Specifically, *ADRA2B* appears in 6 of the top 25 prioritized gene combinations. Other obvious candidate disease genes, including *CTNND2* and *SEMA5A*, located within the FAME3 locus, did not appear in our gene prioritization list. In addition, candidate genes *HTR3D* and *KCNMB3*, located within the FAME4 locus, were prioritized highly in the list and multiple times. These genes encode ion channel receptors that regulate neuron excitability, representing promising candidate disease genes (Niesler et al. 2003; Uebele et al. 2000). Of the genes that were prioritized within the FAME1 and FAME3 loci, not much is known about their function and in-depth analysis of sequencing data may not have been performed.

Discussion

Our analysis demonstrates how difficult the FAME2 critical region is to analyze, even for the IBD analysis. FAME2 overlaps the centromere of chromosome 2, which is a region that is difficult to sequence and hence not as well represented on the SNP genotyping microarrays. We experienced difficulties with insufficient marker information over this region that limited our ability to refine the FAME2 critical region further.

Considering genes within the FAME1, FAME2, FAME3, and FAME4 critical regions, we identified several candidate genes using an in silico gene prioritization method utilizing brain gene expression data, under the assumption that the causal genes in the four loci are likely to be part of the same pathway. This is a reasonable assumption given that the FAME1, FAME2, FAME3, and FAME4 families show a distinctively shared phenotype. As opposed to *ADRA2B*, a receptor that regulates neurotransmitter release, prioritized candidate genes *KCNIP3*, *HTR3D*, and *KCNMB3* are ion channel receptors that regulate neuron excitability, suggesting a common function of potential disease genes from multiple loci. Interestingly, *HTR3D* regulates neuron excitability in response to serotonin; a neurotransmitter associated with neurological disorders, including epilepsy (Wada et al. 1997). Obvious candidate genes from the FAME1 and FAME3 loci were not prioritized using this approach, suggesting that perhaps, the less-obvious candidate disease

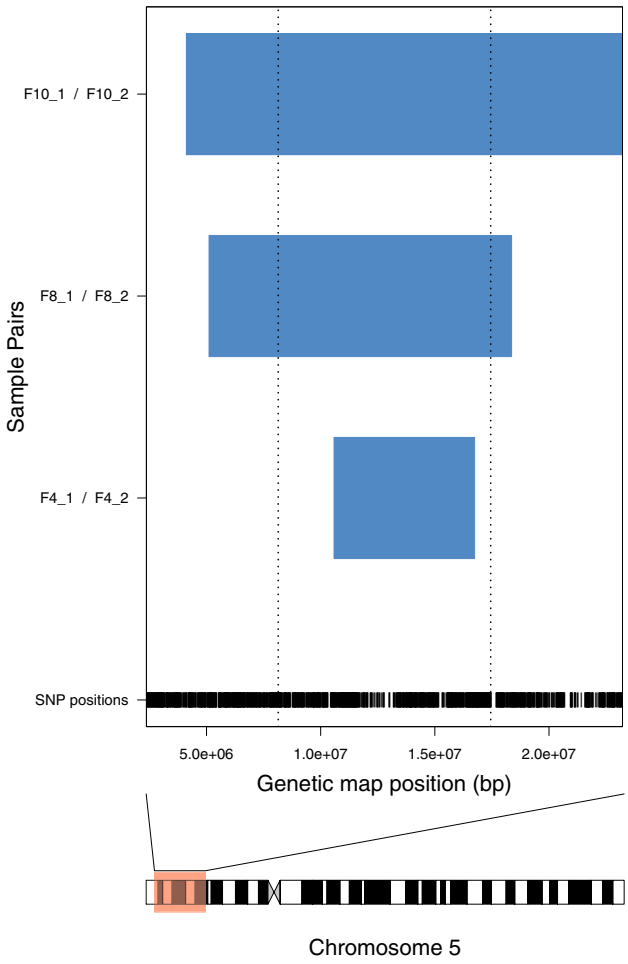


Fig. 4 IBD results across the FAME3 locus, drawn using R. The *y-axis* shows the pair identifiers, and the *x-axis* displays the genetic map position in base pairs along chromosome 5 using hg19. *Solid blue rectangles* are regions, where pairs share one allele IBD, while *checkered yellow rectangles* represent two alleles-shared IBD (none detected). The *dotted vertical lines* mark the FAME3 linkage region as in Depienne et al. (2010). Positions of SNPs included in the analysis can be seen below the IBD tracts along with an ideogram highlighting the interval plotted. All parent-offspring IBD segments have been excluded from the figure; specifically, IBD within the pairs from families *1*, *3*, *9* and *11* (colour figure online)

genes with unknown functions should be scrutinized more deeply for causal variants from these regions.

Our data show that there are multiple founders for each of the large families mapped to the FAME2 locus, suggesting that there may be different causal variants in each of families. Importantly, these observations reduce the chance that FAME2 is due to a common polymorphism that would be filtered out by a typical exome analysis. Variant assessment on candidate FAME genes has not included analysis of less commonly detected polymorphisms, such as microrearrangements and variants in non-coding regions in many families (Guerrini et al. 2001; Saint-Martin et al. 2008; Depienne et al. 2010). Our data suggest that the mutations causing FAME may implicate genes that have previously not been considered and may require the detection of non-coding variants or unusual mutations that are difficult to identify with current short-read-based sequencing methods, such as simple repeat expansions.

Acknowledgments We thank the families for their participation in this study. This work was supported by the National Health and Medical Research Council (NHMRC) Program Grant (628952) to J.G. M.B. was supported by an NHMRC Senior Research Fellowship (1002098) and NHMRC Program Grant (APP1054618). L.H. was supported by The John and Patricia Farrant Scholarship and the Australian Postgraduate Award Scholarship. This work was also supported by Victorian State Government Operational Infrastructure Support, Australian Government NHMRC IRIISS funding, Fondation Maladies rares, Assistance publique des hôpitaux de Paris (AP-HP) and Université Pierre et Marie-Curie (UPMC).

Compliance with ethical standards

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

References

Aerts S, Lambrechts D, Maity S et al (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24:537–544. doi:10.1038/nbt1203

Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet Epidemiol 33:266–274. doi:10.1002/gepi.20378

Anvar SY, Khachatryan L, Vermaat M et al (2014) Determining the quality and complexity of next-generation sequencing data without a reference genome. Genome Biol 15:555. doi:10.1186/s13059-014-0555-3

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN 30:107–117. doi:10.1016/S0169-7552(98)00110-X

Brown MD, Glazner CG, Zheng C, Thompson EA (2012) Inferring coancestry in population samples in the presence of linkage disequilibrium. Genetics 190(4):1447–1460

Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86:526–539. doi:10.1016/j.ajhg.2010.02.021

Choo KHA (1998) Why is the centromere so cold? Genome Res 8:81–82

Crompton DE, Sadleir LG, Bromhead CJ et al (2012) Familial adult myoclonic epilepsy: recognition of mild phenotypes and refinement of the 2q locus. Arch Neurol-Chicago 69:474–481. doi:10.1001/archneurol.2011.584

De Fusco M, Vago R, Striano P et al (2014) The α2B-adrenergic receptor is mutant in cortical myoclonus and epilepsy. Ann Neurol 75:77–87. doi:10.1002/ana.24028

Depienne C, Magnin E, Bouteiller D et al (2010) Familial cortical myoclonic tremor with epilepsy: the third locus (FCMTE3)

maps to 5p. Neurology 74:2000–2003. doi:10.1212/WNL.0b013e3181e396a8

Elia M, Musumeci SA, Ferri R et al (1998) Familial cortical tremor, epilepsy, and mental retardation: a distinct clinical entity? Arch Neurol 55:1569–1573

Freytag S, Gagnon-Bartsch J, Speed TP, Bahlo M (2015) Systematic noise degrades gene co-expression signals but can be corrected. BMC Bioinf 16:309. doi:10.1186/s12859-015-0745-3

Guerrini R, Bonanni P, Patrignani A et al (2001) Autosomal dominant cortical myoclonus and epilepsy (ADCME) with complex partial and generalized seizures: a newly recognized epilepsy syndrome with linkage to chromosome 2p11.1-q12.2. Brain 124:2459–2475

Henden L, Wakeham D, Bahlo M (2016) XIBD: software for inferring pairwise identity by descent on the X chromosome. Bioinformatics. doi:10.1093/bioinformatics/btw124

Ikeda A, Kakigi R, Funai N et al (1990) Cortical tremor: a variant of cortical reflex myoclonus. Neurology 40(10):1561–1565

Kang HJ, Kawasawa YI, Cheng F et al (2011) Spatio-temporal transcriptome of the human brain. Nature 478:483–489. doi:10.1038/nature10523

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. Cell 155:27–38. doi:10.1016/j.cell.2013.09.006

Licchetta L, Pippucci T, Bisulli F et al (2013) A novel pedigree with familial cortical myoclonic tremor and epilepsy (FCMTE): clinical characterization, refinement of the FCMTE2 locus, and confirmation of a founder haplotype. Epilepsia 54:298–1306. doi:10.1111/epi.12216

Madia F, Striano P, Di Bonaventura C et al (2008) Benign adult familial myoclonic epilepsy (BAFME): evidence of an extended founder haplotype on chromosome 2p11.1-q12.2 in five Italian families. Neurogenetics 9:139–142. doi:10.1007/s10048-008-0118-4

Mikami M, Yasuda T, Terao A et al (1999) Localization of a gene for benign adult familial myoclonic epilepsy to chromosome 8q23.3-q24.1. Am J Hum Genet 65:745–751

Mori S, Nakamura M, Yasuda T, Ueno SI, Kaneko S, Sano A (2011) Remapping and mutation analysis of benign adult familial myoclonic epilepsy in a Japanese pedigree. J Hum Genet 56:742–747. doi:10.1038/jhg.2011.93

Nava C, Keren B, Mignot C et al (2014) Prospective diagnostic analysis of copy number variants using SNP microarrays in individuals with autism spectrum disorders. Euro J Hum Genet 22:71–78. doi:10.1038/ejhg.2013.88

Niesler B, Frank B, Kapeller J, Rappold GA (2003) Cloning, physical mapping and expression analysis of the human 5-HT_3 serotonin receptor-like genes *HTR3C*, *HTR3D* and *HTR3E*. Gene 310:101–111. doi:10.1016/S0378-1119(03)00503-1

Oliver KL, Lukic V, Thorne NP, Berkovic SF, Scheffer IE, Bahlo M (2014) Harnessing gene expression networks to prioritize candidate epileptic encephalopathy genes. PLoS One 9:e102079. doi:10.1371/journal.pone.0102079

Plaster NM, Uyama E, Uchino M, Ikeda T, Flanigan KM, Kondo I, Ptácek LJ (1999) Genetic localization of the familial adult myoclonic epilepsy (FAME) gene to chromosome 8q24. Neurology 53:1180–1183

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. IEEE 77:257–286

Saint-Martin C, Bouteiller D, Stevanin G et al (2008) Refinement of the 2p11.1-q12.2 locus responsible for cortical tremor associated with epilepsy and exclusion of candidate genes. Neurogenetics 9:69–71. doi:10.1007/s10048-007-0107-z

Stogmann E, Reinthaler E, Eltawil S et al (2013) Autosomal recessive cortical myoclonic tremor and epilepsy: association with a mutation in the potassium channel associated gene CNTN2. Brain 136:1155–1160. doi:10.1093/brain/awt068

Striano P, Madia F, Minetti C, Striano S, Zara F (2005) Electroclinical and genetic findings in a family with cortical tremor, myoclonus, and epilepsy. Epilepsia 46:1993–1995. doi:10.1111/j.1528-1167.2005.00346.x

The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58

Uebele VN, Lagrutta A, Wade T et al (2000) Cloning and functional expression of two families of β-subunits of the large conductance calcium-activated K^+ channel. J Biol Chem 375:23211–23218. doi:10.1074/jbc.M910187199

Wada Y, Shiraishi J, Nakamura M, Koshino Y (1997) Effects of the 5-HT3 receptor agonist 1-(m-chlorophenyl)-biguanide in the rat kindling model of epilepsy. Brain Res 759(2):313–316. doi:10.1016/S0006-8993(97)00366-1

Yeetong P, Ausavarat S, Bhidayasiri R et al (2013) A newly identified locus for benign adult familial myoclonic epilepsy on chromosome 3q26.32-3q28. Eur J Hum Genet 21:225–228. doi:10.1038/ejhg.2012.133

4.2 Supplementary tables

Supplementary Table 1 Study design underlying the four large microarray gene expression datasets. PCW stands for post conception weeks

Study	Age Range (youngest, oldest)	Platform	No. Brains	No. Samples	Approx. No. Genes
Hawrylycz et al	(24 years, 57 years)	Agilent 64K (custom array)	10	3,546	20,000
Trabzuni et al	(16 years, 102 years)	Affymetrix Human Exon 1.0ST 8x60K	134	1,222	17,500
Kang et al	(4 PCW, 82 years)	Affymetrix Human Exon 1.0 ST	57	1,329	17,500
Colantuoni et al	(14 PCW, 80 years)	Illumina (custom array)	266	266	20,000
Hernandez et al	(0.6 years, 102 years)	Illumina HumanHT-12 V3.0	397	908	18,000

Supplementary Table 2 List of 421 known epilepsy genes used in gene prioritization

ABCB1, ABCC1, ABCC2, ABCC5, ABCG2, ACMSD, ACOT7, ADAM22, ADIPOQ, ADK, ADORA1, ADSL, ALB, ALDH5A1, ALDH7A1, ALG13, ALPL, ALX4, ANXA7, AP3M2, AP4E1, APLN, APOE, APP, AQP1, AQP4, ARF6, ARFGEF2, ARHGAP11B, ARHGEF9, ARPC2, ARX, ASAH1, ATF3, ATN1, ATP1A2, ATP1A3, ATP6V0C, ATXN10, BCKDHA, BCS1L, BDNF, BRD2, C10orf2, C3, CACNA1A, CACNA1E, CACNA1G, CACNA1H, CACNB4, CACNG3, CALB2, CALHM1, CAPN1, CASP2, CASP3, CASR, CCL2, CCL4, CCM2, CCR5, CD40, CD40LG, CDKL5, CHD2, CHL1, CHRFAM7A, CHRNA2, CHRNA4, CHRNA7, CHRNB2, CLCN2, CLN3, CLN5, CLN6, CLN8, CNKSR2, CNN3, CNTN2, CNTNAP2, COL4A1, COL6A2, COX10, COX15, CPA6, CRH, CRMP1, CSMD3, CST3, CSTB, CTSD, CTSF, CXCL8, CYP2C19, CYP2C9, CYP3A4, CYP3A5, D2HGDH, DAPK1, DARS2, DBH, DBP, DCX, DEPDC5, DFFB, DLG2, DNAJC5, DNAJC6, DNM1, DNMT1, DNMT3A, DPYSL2, DSCAM, DTNBP1, DYRK1A, EFHC1, EFHC2, EGFR, EGR1, EHMT1, EIF2S1, ELP4, EMP1, EMX2, ENO2, EPHX1, EPM2A, EPM2AIP1, ERMN, ERN1, FAM3C, FGFR3, FLNA, FOXG1, FOXRED1, GABARAP, GABBR1, GABBR2, GABRA1, GABRA6, GABRB1, GABRB2, GABRB3, GABRD, GABRE, GABRG2, GABRR2, GAD1, GAD2, GBA, GFAP, GHRL, GJD2, GLI3, GLUD1, GLUL, GNAO1, GNB3, GOSR2, GPHN, GPR56, GRIA1, GRIA2, GRIA3, GRIK1, GRIN1, GRIN2A, GRIN2B, GRM1, GRM5, GRN, GSN, GSTA4, GSTP1, HCN1, HCN2, HCN3, HCN4, HCRT, HDAC2, HEPACAM, HIP1, HLA-B, HLA-DQA1, HLA-DQB1, HNRNPU, HP, HSPA8, HSPB1, HSPBAP1, HTR1A, HTR7, IDH1, IER3IP1, IGF1, IL1A, IL1B, IL1RN, IL4, IL6, INSR, ITGA2, JRK, KCNA1, KCNA2, KCNAB1, KCNB1, KCNC1, KCND1, KCND2, KCNH2, KCNJ10, KCNJ11, KCNJ3, KCNK3, KCNK9, KCNMA1, KCNMB1, KCNMB2, KCNMB3, KCNMB4, KCNQ1, KCNQ2, KCNQ3, KCNT1, KCNV2, KCTD7, KIF5A, KL, KRIT1, L2HGDH, LAMB1, LAMC3, LEPR, LGI1, LGI2, LGI4,

LIAS, MAP2, MAPT, MBD5, MC3R, MDM2, ME2, MECP2, MED1, MEF2C, MFSD8, MICAL1, MLC1, MMP9, MRI1, MSX2, MT2A, MTHFR, MTOR, MVP, NCAM1, NDUFA1, NDUFA10, NDUFA12, NDUFA2, NDUFA9, NDUFAF2, NDUFS3, NDUFS4, NDUFS7, NDUFS8, NEDD4L, NFKB1, NHLRC1, NIPA1, NIPA2, NOS1, NPPB, NR1I2, NRG1, NTNG1, NTNG2, NUCB2, OLIG2, OPRM1, OTX1, P2RY1, P2RY2, P2RY4, PAFAH1B1, PANX1, PANX2, PCDH19, PDXK, PDYN, PGF, PHF6, PHLDA1, PHOX2A, PHOX2B, PLAUR, PLCB1, PNKP, PNOC, PNPO, POLG, PPFIA1, PPP1R3D, PPT1, PRICKLE1, PRICKLE2, PRNP, PRODH, PROM1, PRRT2, PSEN1, PTGS2, PTPRD, QARS, QPRT, RALBP1, RASGRF1, RBFOX1, RBFOX3, RCN2, RELN, REST, RHOA, RNF115, RORA, RTN4, S100B, SAA1, SCARB2, SCN1A, SCN1B, SCN2A, SCN2B, SCN3A, SCN3B, SCN5A, SCN8A, SCN9A, SDHA, SERPINI1, SEZ6, SGK1, SHH, SLC12A5, SLC13A5, SLC16A1, SLC16A7, SLC18A2, SLC1A1, SLC1A2, SLC1A3, SLC25A22, SLC2A1, SLC35A2, SLC35A3, SLC6A3, SLC6A4, SLC6A8, SLIT2, SNAP25, SNIP1, SNX25, SOD1, SOD2, SPTAN1, SRPX2, SSTR2, ST3GAL3, ST3GAL5, STAMBP, STIM1, STIM2, STMN1, STRADA, STX1A, STX1B, STXBP1, SUCLA2, SURF1, SV2A, SYN1, SYN2, SYNGAP1, SYT1, SYT11, SZT2, TACR1, TAP1, TBC1D24, TK2, TLN2, TLR4, TNF, TNK2, TP53, TPP1, TRAPPC10, TRMT44, TRPC4, TRPM2, TRPV1, TSC1, TSC2, TSEN2, TSEN34, TSEN54, TSPEAR, TSPO, TUBA1A, UBA1, UBC, UBE3A, UCP2, UGT1A4, VAMP2, VDR, WASL, WNT8B, WWOX

Supplementary Table 3 Parameter estimates for all pairs of individuals. Individual 1 from family 1 is denoted 'F1_1'. The meiosis estimates the total number of meiosis separating the pair. Meiosis of 14 corresponds to unrelated individuals. The last 3 columns are the probability of sharing 0, 1 and 2 alleles IBD respectively. All parameter estimates were calculated using formula in Purcell et al. (2007)

Individual 1	Individual 2	Meiosis	Pr(IBD=0)	Pr(IBD=1)	Pr(IBD=2)
F1_1	F1_2	2.01	0.001	0.996	0.003
F1_1	F3_1	8.16	0.993	0.007	0
F1_1	F3_2	14	1	0	0
F1_2	F3_1	14	1	0	0
F1_2	F3_2	14	1	0	0
F1_1	F4_1	14	1	0	0
F1_1	F4_2	14	1	0	0
F1_2	F4_1	14	1	0	0
F1_2	F4_2	14	1	0	0
F1_1	F5_1	6.8	0.982	0.018	0
F1_1	F5_2	14	1	0	0
F1_2	F5_1	14	1	0	0
F1_2	F5_2	14	1	0	0
F1_1	F6_1	14	1	0	0

F1_1	F6_2	14	1	0	0
F1_2	F6_1	14	1	0	0
F1_2	F6_2	14	1	0	0
F1_1	F8_1	14	1	0	0
F1_1	F8_2	14	1	0	0
F1_2	F8_1	14	1	0	0
F1_2	F8_2	14	1	0	0
F1_1	F9_1	14	1	0	0
F1_1	F9_2	14	1	0	0
F1_2	F9_1	14	1	0	0
F1_2	F9_2	14	1	0	0
F1_1	F10_1	6.97	0.984	0.016	0
F1_1	F10_2	14	1	0	0
F1_2	F10_1	7.38	0.988	0.012	0
F1_2	F10_2	14	1	0	0
F1_1	F11_1	14	1	0	0
F1_1	F11_2	14	1	0	0
F1_2	F11_1	14	1	0	0
F1_2	F11_2	14	1	0	0
F2_1	F1_1	14	1	0	0
F2_1	F1_2	14	1	0	0
F2_2	F1_1	14	1	0	0
F2_2	F1_2	14	1	0	0
F2_1	F2_2	14	1	0	0
F2_1	F3_1	14	1	0	0
F2_1	F3_2	14	1	0	0
F2_2	F3_1	14	1	0	0
F2_2	F3_2	14	1	0	0
F2_1	F4_1	14	1	0	0
F2_1	F4_2	14	1	0	0
F2_2	F4_1	14	1	0	0
F2_2	F4_2	14	1	0	0
F2_1	F5_1	14	1	0	0
F2_1	F5_2	14	1	0	0
F2_2	F5_1	7.64	0.99	0.01	0
F2_2	F5_2	14	1	0	0
F2_1	F6_1	14	1	0	0
F2_1	F6_2	14	1	0	0
F2_2	F6_1	14	1	0	0
F2_2	F6_2	14	1	0	0
F2_1	F7_1	14	1	0	0
F2_1	F7_2	14	1	0	0
F2_2	F7_1	14	1	0	0

F2_2	F7_2	14	1	0	0
F2_1	F8_1	14	1	0	0
F2_1	F8_2	14	1	0	0
F2_2	F8_1	14	1	0	0
F2_2	F8_2	14	1	0	0
F2_1	F9_1	14	1	0	0
F2_1	F9_2	14	1	0	0
F2_2	F9_1	14	1	0	0
F2_2	F9_2	14	1	0	0
F2_1	F10_1	14	1	0	0
F2_1	F10_2	14	1	0	0
F2_2	F10_1	14	1	0	0
F2_2	F10_2	14	1	0	0
F2_1	F11_1	14	1	0	0
F2_1	F11_2	14	1	0	0
F2_2	F11_1	14	1	0	0
F2_2	F11_2	14	1	0	0
F3_1	F3_2	2.01	0.001	0.996	0.004
F3_1	F4_1	14	1	0	0
F3_1	F4_2	6.88	0.983	0.017	0
F3_2	F4_1	7.06	0.985	0.015	0
F3_2	F4_2	6.88	0.983	0.017	0
F3_1	F5_1	7.27	0.987	0.013	0
F3_1	F5_2	14	1	0	0
F3_2	F5_1	14	1	0	0
F3_2	F5_2	14	1	0	0
F3_1	F6_1	8.97	0.996	0.004	0
F3_1	F6_2	14	1	0	0
F3_2	F6_1	14	1	0	0
F3_2	F6_2	14	1	0	0
F3_1	F8_1	14	1	0	0
F3_1	F8_2	14	1	0	0
F3_2	F8_1	14	1	0	0
F3_2	F8_2	14	1	0	0
F3_1	F9_1	14	1	0	0
F3_1	F9_2	14	1	0	0
F3_2	F9_1	14	1	0	0
F3_2	F9_2	14	1	0	0
F3_1	F10_1	9.97	0.998	0.002	0
F3_1	F10_2	8.38	0.994	0.006	0
F3_2	F10_1	14	1	0	0
F3_2	F10_2	14	1	0	0
F3_1	F11_1	14	1	0	0

F3_1	F11_2	14	1	0	0
F3_2	F11_1	14	1	0	0
F3_2	F11_2	14	1	0	0
F4_1	F4_2	4.92	0.863	0.132	0.005
F4_1	F5_1	14	1	0	0
F4_1	F5_2	14	1	0	0
F4_2	F5_1	7.27	0.987	0.013	0
F4_2	F5_2	14	1	0	0
F4_1	F6_1	14	1	0	0
F4_1	F6_2	14	1	0	0
F4_2	F6_1	14	1	0	0
F4_2	F6_2	14	1	0	0
F4_1	F8_1	14	1	0	0
F4_1	F8_2	14	1	0	0
F4_2	F8_1	14	1	0	0
F4_2	F8_2	14	1	0	0
F4_1	F9_1	14	1	0	0
F4_1	F9_2	14	1	0	0
F4_2	F9_1	14	1	0	0
F4_2	F9_2	14	1	0	0
F4_1	F10_1	6.8	0.982	0.018	0
F4_1	F10_2	14	1	0	0
F4_2	F10_1	14	1	0	0
F4_2	F10_2	14	1	0	0
F4_1	F11_1	14	1	0	0
F4_1	F11_2	14	1	0	0
F4_2	F11_1	14	1	0	0
F4_2	F11_2	14	1	0	0
F5_1	F5_2	3.07	0.524	0.475	0.001
F5_1	F6_1	14	1	0	0
F5_1	F6_2	8.97	0.996	0.004	0
F5_2	F6_1	14	1	0	0
F5_2	F6_2	7.97	0.992	0.008	0
F5_1	F8_1	14	1	0	0
F5_1	F8_2	7.8	0.991	0.009	0
F5_2	F8_1	14	1	0	0
F5_2	F8_2	14	1	0	0
F5_1	F9_1	14	1	0	0
F5_1	F9_2	8.16	0.993	0.007	0
F5_2	F9_1	14	1	0	0
F5_2	F9_2	14	1	0	0
F5_1	F10_1	6.51	0.978	0.022	0
F5_1	F10_2	14	1	0	0

F5_2	F10_1	14	1	0	0
F5_2	F10_2	14	1	0	0
F5_1	F11_1	14	1	0	0
F5_1	F11_2	14	1	0	0
F5_2	F11_1	14	1	0	0
F5_2	F11_2	14	1	0	0
F6_1	F6_2	4.59	0.832	0.166	0.002
F6_1	F9_1	14	1	0	0
F6_1	F9_2	14	1	0	0
F6_2	F9_1	14	1	0	0
F6_2	F9_2	14	1	0	0
F6_1	F10_1	8.64	0.995	0.005	0
F6_1	F10_2	14	1	0	0
F6_2	F10_1	6.64	0.98	0.02	0
F6_2	F10_2	14	1	0	0
F6_1	F11_1	14	1	0	0
F6_1	F11_2	14	1	0	0
F6_2	F11_1	14	1	0	0
F6_2	F11_2	14	1	0	0
F7_1	F1_1	14	1	0	0
F7_1	F1_2	14	1	0	0
F7_2	F1_1	14	1	0	0
F7_2	F1_2	14	1	0	0
F7_1	F3_1	14	1	0	0
F7_1	F3_2	14	1	0	0
F7_2	F3_1	14	1	0	0
F7_2	F3_2	14	1	0	0
F7_1	F4_1	14	1	0	0
F7_1	F4_2	14	1	0	0
F7_2	F4_1	14	1	0	0
F7_2	F4_2	14	1	0	0
F7_1	F5_1	14	1	0	0
F7_1	F5_2	14	1	0	0
F7_2	F5_1	14	1	0	0
F7_2	F5_2	14	1	0	0
F7_1	F6_1	14	1	0	0
F7_1	F6_2	14	1	0	0
F7_2	F6_1	14	1	0	0
F7_2	F6_2	8.38	0.994	0.006	0
F7_1	F7_2	5.01	0.938	0.062	0
F7_1	F8_1	9.38	0.997	0.003	0
F7_1	F8_2	6.88	0.983	0.017	0
F7_2	F8_1	14	1	0	0

F7_2	F8_2	7.97	0.992	0.008	0
F7_1	F9_1	14	1	0	0
F7_1	F9_2	14	1	0	0
F7_2	F9_1	7.64	0.99	0.01	0
F7_2	F9_2	14	1	0	0
F7_1	F10_1	6.01	0.969	0.031	0
F7_1	F10_2	14	1	0	0
F7_2	F10_1	14	1	0	0
F7_2	F10_2	14	1	0	0
F7_1	F11_1	14	1	0	0
F7_1	F11_2	14	1	0	0
F7_2	F11_1	14	1	0	0
F7_2	F11_2	14	1	0	0
F8_1	F6_1	14	1	0	0
F8_1	F6_2	14	1	0	0
F8_2	F6_1	14	1	0	0
F8_2	F6_2	14	1	0	0
F8_1	F8_2	5.52	0.911	0.087	0.002
F8_1	F9_1	14	1	0	0
F8_1	F9_2	14	1	0	0
F8_2	F9_1	14	1	0	0
F8_2	F9_2	14	1	0	0
F8_1	F10_1	5.72	0.962	0.038	0
F8_1	F10_2	14	1	0	0
F8_2	F10_1	6.57	0.979	0.021	0
F8_2	F10_2	14	1	0	0
F8_1	F11_1	14	1	0	0
F8_1	F11_2	14	1	0	0
F8_2	F11_1	14	1	0	0
F8_2	F11_2	14	1	0	0
F9_1	F9_2	1	0	1	0
F9_1	F10_1	7.51	0.989	0.011	0
F9_1	F10_2	14	1	0	0
F9_2	F10_1	6.88	0.983	0.017	0
F9_2	F10_2	6.8	0.982	0.018	0
F9_1	F11_1	8.16	0.993	0.007	0
F9_1	F11_2	14	1	0	0
F9_2	F11_1	7.16	0.986	0.014	0
F9_2	F11_2	14	1	0	0
F10_1	F10_2	3.32	0.8	0.2	0
F10_1	F11_1	14	1	0	0
F10_1	F11_2	14	1	0	0
F10_2	F11_1	14	1	0	0

F10_2	F11_2	14	1	0	0
F11_1	F11_2	1	0	1	0

Supplementary Table 4 Description of IBD tracts inferred over the FAME2 critical region. All genetic map positions are from hg19

Ind 1	Ind 2	Chr	Start position (bp)	End position (bp)	No. markers	Length (bp)	Length (cM)	IBD status	Start marker	End marker
F1_1	F1_2	2	72184	243020723	15453	242948539	268.8	1	rs300758	rs4973686
F2_1	F2_2	2	85952162	102205736	538	16253574	7.45	1	rs4832005	rs7603851
F2_1	F3_1	2	85952162	100615241	394	14663079	6.22	1	rs4832005	rs17437101
F2_1	F3_2	2	85995648	100403354	377	14407706	5.92	1	rs4832198	rs2115601
F2_2	F3_1	2	84905096	100679024	494	15773928	7.41	1	rs11126974	rs11681737
F2_2	F3_2	2	84921433	100665673	491	15744240	7.39	1	rs1192295	rs2028137
F2_1	F4_1	2	85995648	101337982	438	15342334	6.43	1	rs4832198	rs11123842
F2_1	F4_2	2	85976645	100421767	382	14445122	6.07	1	rs1465823	rs1568786
F2_2	F4_1	2	78968942	100474071	825	21505129	12.37	1	rs4853406	rs17023232
F2_2	F4_2	2	79225690	100403354	803	21177664	11.85	1	rs402221	rs2115601
F2_1	F5_1	2	85976645	104498739	660	18522094	9.02	1	rs1465823	rs6717408
F2_1	F5_2	2	87852863	103950575	564	16097712	7.62	1	rs10180746	rs10191917
F2_2	F5_1	2	85654299	102421978	581	16767679	7.9	1	rs1877954	rs6732726
F2_2	F5_2	2	89052026	102604649	398	13552623	4.09	1	rs335124	rs12467316
F3_1	F3_2	2	72184	243020723	15453	242948539	268.8	1	rs300758	rs4973686
F3_1	F4_1	2	84905096	106053343	856	21148247	12.04	1	rs11126974	rs7583367
F3_1	F4_2	2	84800898	106045613	860	21244715	12	1	rs13002679	rs7594621
F3_2	F4_1	2	84921433	106053343	855	21131910	12.03	1	rs1192295	rs7583367
F3_2	F4_2	2	84921433	106053343	855	21131910	12.03	1	rs1192295	rs7583367
F3_1	F5_1	2	85731858	101343140	462	15611282	6.94	1	rs6750610	rs2137671
F3_1	F5_2	2	86627316	100615241	336	13987925	5.66	1	rs1105865	rs17437101
F3_2	F5_1	2	85628983	100403354	410	14774371	6.41	1	rs2229668	rs2115601
F3_2	F5_2	2	86843488	100403354	314	13559866	5.5	1	rs308903	rs2115601
F4_1	F4_2	2	36879521	106511846	4478	69632325	63.25	1	rs10167726	rs933793
F4_1	F5_1	2	85731858	101378095	467	15646237	6.96	1	rs6750610	rs6740105

F4_1	F5_2	2	86364353	101337982	407	14973629	6.15	1	rs2241434	rs11123842
F4_2	F5_1	2	85731858	100232743	387	14500885	6.25	1	rs6750610	rs12616127
F4_2	F5_2	2	87852863	100421767	301	12568904	4.89	1	rs10180746	rs1568786
F5_1	F5_2	2	88295232	189122229	5495	100826997	83.63	1	rs1441649	rs7582137
F6_1	F6_2	2	75390741	101480885	1131	26090144	17.23	1	rs17010840	rs2043534
F7_1	F7_2	2	86013029	101318308	432	15305279	6.39	1	rs13386681	rs2942883
F8_1	F8_2	2	56029444	139104124	4638	83074680	74.04	1	rs6712017	rs4550720
F9_1	F9_2	2	72184	243020723	15453	242948539	268.8	1	rs300758	rs4973686
F11_1	F11_2	2	72184	151622783	9819	151550599	167.9	1	rs300758	rs1519756

Supplementary Table 5 Description of IBD tracts inferred over the FAME3 critical region. All genetic map positions are from hg19

Ind 1	Ind 2	Chr	Start position (bp)	End position (bp)	No. markers	Length (bp)	Length (cM)	IBD status	Start marker	End marker
F1_1	F1_2	5	38139	180682862	11879	180644723	204.04	1	rs10076494	rs2545093
F3_1	F3_2	5	38139	180682862	11879	180644723	204.04	1	rs10076494	rs2545093
F4_1	F4_2	5	10585111	16741099	479	6155988	7.44	1	rs7712927	rs10051930
F8_1	F8_2	5	5110692	18362647	1223	13251955	22.26	1	rs814790	rs2950483
F9_1	F9_2	5	38139	180682862	11879	180644723	204.04	1	rs10076494	rs2545093
F10_1	F10_2	5	4117568	34700704	2483	30583136	42.73	1	rs11134032	rs11740818
F11_1	F11_2	5	38139	180682862	11879	180644723	204.04	1	rs10076494	rs2545093

Supplementary Table 6 RefSeq genes overlapping the FAME1, FAME2, FAME3 and FAME4 critical regions. Genes coloured red were found in at least one gene-expression dataset

FAME1 genes	FAME2 genes	FAME3 genes	FAME4 genes
AARD	*ACTR1B*	*ADCY2*	*ABCC5*
ANXA13	*ACTR3BP2*	*ANKH*	*ABCC5-AS1*
ATAD2	*ADRA2B*	*ANKRD33B*	*ABCF3*
C8orf76	*ANKRD20A8P*	*BASP1*	*ACTL6A*
COL14A1	*ANKRD23*	*C5orf49*	*ADIPOQ*
COLEC10	*ANKRD36*	*CCT5*	*ADIPOQ-AS1*
DEPTOR	*ANKRD36B*	*CMBL*	*AHSG*
DERL1	*ANKRD39*	*CTD-2201E9.1*	*ALG3*
DSCC1	*ARID5A*	*CTD-2350J17.1*	*AP2M1*
EIF3H	*ASTL*	*CTNND2*	*ATP11B*
ENPP2	*CIAO1*	*DAP*	*B3GNT5*
EXT1	*CNNM3*	*DNAH5*	*BCL6*
FAM83A	*CNNM4*	*FAM105A*	*C3orf70*
FAM83A-AS1	*COX5B*	*FAM134B*	*CCDC39*
FBXO32	*DUSP2*	*FAM173B*	*CCDC50*
HAS2	*FAHD2A*	*FASTKD3*	*CHRD*
HAS2-AS1	*FAHD2B*	*FBXL7*	*CLCN2*
KLHL38	*FAHD2CP*	*FLJ33360*	*CLDN1*
LINC01151	*FAM178B*	*LINC01018*	*CLDN16*
LOC101927543	*FAM95A*	*LINC01194*	*CRYGS*
LOC105375734	*FER1L5*	*LOC100120744*	*DCUN1D1*
MAL2	*GGT8P*	*LOC100505625*	*DGKG*
MED30	*GPAT2*	*LOC101929284*	*DNAJB11*
MIR3610	*ITPRIPL1*	*LOC101929412*	*DNAJC19*
MIR4663	*KANSL3*	*LOC101929454*	*DVL3*
MRPL13	*KCNIP3*	*LOC101929505*	*ECE2*
MTBP	*LINC00342*	*LOC101929524*	*EHHADH*
NOV	*LINC01125*	*LOC285692*	*EHHADH-AS1*
RAD21	*LMAN2L*	*LOC285696*	*EIF2B5*
RAD21-AS1	*LOC100506076*	*LOC401177*	*EIF2B5-AS1*
SAMD12	*LOC100506123*	*LOC442132*	*EIF4A2*
SAMD12-AS1	*LOC10192703*	*LOC729506*	*EIF4G1*
SLC30A8	*LOC442028*	*MARCH6*	*EPHB3*
SNTB1	*LOC654342*	*MARCH11*	*ETV5*
TAF2	*MAL*	*MED10*	*FAM131A*
TBC1D31	*MIR3127*	*MIR4278*	*FETUB*
TNFRSF11B	*MRPS5*	*MIR4454*	*FGF12*

UTP23	*NCAPH*	*MIR4454*	*FGF12-AS1*
WDYHV1	*NEURL3*	*MIR4458*	*FLJ42393*
ZHX1	*PROM2*	*MIR4636*	*FLJ46066*
ZHX1-C8orf76	*SEMA4C*	*MIR4637*	*FXR1*
ZHX2	*SNRNP200*	*MIR6131*	*GMNC*
	STARD7	*MIR887*	*GNB4*
	STARD7-AS1	*MTRR*	*HRG*
	TEKT4	*MYO10*	*HTR3D*
	TMEM127	*NSUN2*	*HTR3E*
	TMEM131	*OTULIN*	*HTR3E-AS1*
	TRIM43	*PAPD7*	*IDF2BP2-AS1*
	TRIM43B	*ROPN1L*	*IGF2BP2*
	VWA3B	*ROPN1L-AS1*	*IL1RAP*
	ZAP70	*SEMA5A*	*KCCAT211*
	ZNF2	*SNHG18*	*KCNMB2*
	ZNF514	*SNORD123*	*KCNMB2-AS1*
		SRD5A1	*KCNMB3*
		TAS2R1	*KLHL24*
		TRIO	*KLHL6*
		UBE2QL1	*KLHL6-AS1*
		ZNF622	*KNG1*
			LAMP3
			LINC00501
			LINC00578
			LINC00888
			LINC01014
			LINC01206
			LINC01208
			LINC01209
			LINCR-002
			LIPH
			LOC100131635
			LOC100505609
			LOC101928739
			LOC101928882
			LOC101928992
			LOC101929106
			LOC102724604
			LOC102724699
			LOC105374244
			LOC105374250
			LOC105374266
			LOC253573

			LOC344887 *LPP* *LPP-AS1* *LPP-AS2* *MAGEF1* *MAP3K13* *MAP6D1* *MCCC1* *MCF2L2* *MFN1* *MIR1224* *MIR1248* *MIR28* *MIR4448* *MIR548AQ* *MIR5588* *MIR7977* *MIR944* *MRPL47* *NDVFB5* *OSTN* *OSTN-AS1* *P3H2* *P3H2-AS1* *PARL* *PEX5L* *PEX5L-AS2* *PIK3CA* *POLR2H* *PSMD2* *PYDC2* *RFC4* *RNU6-1* *RNU6-2* *RNU6-7* *RNU6-8* *RNU6-9* *RPL39L* *RTP2* *RTP4* *SENP2* *SNAR-I* *SNORA4*

			SNORA63 *SNORA81* *SNORD2* *SNORD66* *SOX2* *SOX2-OT* *SST* *ST6GAL1* *TBCCD1* *TBL1XR1* *THPO* *TMEM207* *TMEM41A* *TP63* *TPRG1* *TPRG1-AS1* *TPRG1-AS2* *TRA2B* *TTC14* *USP13* *UTS2B* *VPS8* *VWA5B2* *YEATS2* *ZMAT3* *ZNF639*

Supplementary Table 7 Gene prioritization results for the FAME1, FAME2, FAME3 and FAME4 loci; ordered by the sum of the page rank

FAME1 genes	FAME2 genes	FAME3 genes	FAME4 genes	Combined Page Rank
AARD	*GPAT2*	*ANKRD33B*	*HTR3D*	0.013296541
KLHL38	*TRIM43B*	*TAS2R1*	*TP63*	0.009211528
KLHL38	*TRIM43B*	*TAS2R1*	*KLHL6*	0.009104841
KLHL38	*TRIM43B*	*TAS2R1*	*HRG*	0.009034682
KLHL38	*TRIM43B*	*TAS2R1*	*KCNMB3*	0.008910074
KLHL38	*TRIM43B*	*TAS2R1*	*EHHADH*	0.008910074
KLHL38	*TRIM43B*	*TAS2R1*	*FETUB*	0.00890013
KLHL38	*TRIM43B*	*TAS2R1*	*TBCCD1*	0.008708937
AARD	*ASTL*	*ANKRD33B*	*HTR3D*	0.008231326
KLHL38	*NCAPH*	*TAS2R1*	*TP63*	0.007722105
KLHL38	*NCAPH*	*TAS2R1*	*FETUB*	0.007617178

KLHL38	*NCAPH*	*TAS2R1*	*KLHL6*	0.007601198
KLHL38	*NCAPH*	*TAS2R1*	*HRG*	0.007509636
KLHL38	*NCAPH*	*TAS2R1*	*KCNMB3*	0.007391684
KLHL38	*NCAPH*	*TAS2R1*	*EHHADH*	0.007391684
KLHL38	*ADRA2B*	*TAS2R1*	*TP63*	0.007302448
AARD	*GGT8P*	*ANKRD33B*	*HTR3D*	0.007220685
KLHL38	*ADRA2B*	*TAS2R1*	*FETUB*	0.00719668
KLHL38	*ADRA2B*	*TAS2R1*	*KLHL6*	0.007180214
KLHL38	*ADRA2B*	*TAS2R1*	*HRG*	0.007088137
KLHL38	*TRIM43*	*TAS2R1*	*TP63*	0.007002472
FAM83A	*TRIM43B*	*TAS2R1*	*TP63*	0.006991272
KLHL38	*ADRA2B*	*TAS2R1*	*KCNMB3*	0.006968735
KLHL38	*ADRA2B*	*TAS2R1*	*EHHADH*	0.006968735
KLHL38	*TRIM43*	*TAS2R1*	*FETUB*	0.006891715
KLHL38	*TRIM43*	*TAS2R1*	*KLHL6*	0.006875687
FAM83A	*TRIM43B*	*TAS2R1*	*KLHL6*	0.006868785
KLHL38	*TRIM43*	*TAS2R1*	*HRG*	0.006777664
FAM83A	*TRIM43B*	*TAS2R1*	*HRG*	0.006776381
COLEC10	*TRIM43B*	*TAS2R1*	*TP63*	0.006746034
FAM83A	*TRIM43B*	*TAS2R1*	*KCNMB3*	0.006658786
FAM83A	*TRIM43B*	*TAS2R1*	*EHHADH*	0.006658786
FAM83A	*TRIM43B*	*TAS2R1*	*TBCCD1*	0.006658786
KLHL38	*TRIM43*	*TAS2R1*	*KCNMB3*	0.006652532
KLHL38	*TRIM43*	*TAS2R1*	*EHHADH*	0.006652532
SLC30A8	*TRIM43B*	*TAS2R1*	*TP63*	0.006643974
COLEC10	*TRIM43B*	*TAS2R1*	*KLHL6*	0.006620877
COLEC10	*TRIM43B*	*TAS2R1*	*HRG*	0.006526535
ANXA13	*TRIM43B*	*TAS2R1*	*TP63*	0.006526535
SLC30A8	*TRIM43B*	*TAS2R1*	*KLHL6*	0.006517515
SLC30A8	*TRIM43B*	*TAS2R1*	*HRG*	0.006418317
COLEC10	*TRIM43B*	*TAS2R1*	*KCNMB3*	0.006403794
COLEC10	*TRIM43B*	*TAS2R1*	*EHHADH*	0.006403794
COLEC10	*TRIM43B*	*TAS2R1*	*TBCCD1*	0.006403794
HAS2	*TRIM43B*	*TAS2R1*	*TP63*	0.006403794
ANXA13	*TRIM43B*	*TAS2R1*	*KLHL6*	0.006398553
ANXA13	*TRIM43B*	*TAS2R1*	*HRG*	0.006298893
SLC30A8	*TRIM43B*	*TAS2R1*	*KCNMB3*	0.006294698
SLC30A8	*TRIM43B*	*TAS2R1*	*EHHADH*	0.006294698
SLC30A8	*TRIM43B*	*TAS2R1*	*TBCCD1*	0.006294698
HAS2	*TRIM43B*	*TAS2R1*	*KLHL6*	0.006272579
EXT1	*KCNIP3*	*LOC285696*	*FGF12*	0.006216727
ANXA13	*TRIM43B*	*TAS2R1*	*KCNMB3*	0.00617169
ANXA13	*TRIM43B*	*TAS2R1*	*EHHADH*	0.00617169

ANXA13	*TRIM43B*	*TAS2R1*	*TBCCD1*	0.00617169
HAS2	*TRIM43B*	*TAS2R1*	*HRG*	0.00617169
HAS2	*TRIM43B*	*TAS2R1*	*KCNMB3*	0.006047487
HAS2	*TRIM43B*	*TAS2R1*	*EHHADH*	0.006047487
HAS2	*TRIM43B*	*TAS2R1*	*TBCCD1*	0.006047487
EXT1	*KCNIP3*	*LOC285696*	*KCNMB2*	0.005595839
NOV	*KCNIP3*	*BASP1*	*SST*	0.005496796

4.3 Additional methods

4.3.1 Selecting best matched HapMap population

We implemented a goodness-of fit test that compared the genotypes for the FAME cohort to the expected genotypes from selected populations (e.g. Caucasians (CEU), Chinese (CHB), Tuscans (TSI) etc.) in order to get the best matched population allele frequencies for our analysis using the popHetTest from LINKDATAGEN[74,77]. We used the HapMap Phase 3 frequency data from Human Genome version 19 (hg19), which contains allele frequencies for all 11 HapMap populations[73]. Sixteen individuals best matched the TSI population while the remaining six best matched the CEU population. The population allele frequencies are very similar between the TSI and CEU populations so we chose the TSI allele frequencies as representative of all individuals in our analysis. LINKDATAGEN was then used to generate FAME genotypes from HapMap SNPs to be used in the IBD analysis.

4.3.2 Gene prioritization data cleaning

In a first pre-processing step we removed arrays that were obvious outliers. Secondly, we applied adaptive removal of unwanted variation to each dataset individually using the Bioconductor-package RUVcorr[78]. Note that we applied RUV with housekeeping genes as our negative controls (excluding housekeeping genes that were also implicated in epilepsy or potential candidates). Thirdly, all datasets were scaled and centered before combining them into a large dataset.

Chapter 5

An introduction to malaria

5.1 Background

Malaria is an infectious disease that is responsible for an estimated 500,000 deaths and more than 200 million clinical cases, annually[79]. Approximately half of the population live in regions with malaria transmission, predominantly in developing countries in sub-tropical regions of the world (Figure 5.1A)[79]. Malaria is caused by the parasite *Plasmodium*, which is transmitted to humans through the bite of an infected female *Anopheles* mosquito. There are 6 species of *Plasmodium* that infect humans, with the most common being *P. falciparum* and *P. vivax*[80]. The species *P. falciparum* is dominant in Sub-Saharan Africa (Figure 5.1B) and is responsible for nearly all malaria deaths with more than 70% occurring in children under the age of 5[79]. This translates to the death of one child every two minutes. While *P. falciparum* is by far the deadliest *Plasmodium* species that infects humans, *P. vivax*, which is most common outside of Africa (Figure 5.1C), causes the most morbidity as this species can lie dormant in the human host for undefined periods of time and result in relapse infections[79].

Due to the burden of this disease much work has been done to control malaria and a strategy has been implemented to reduce malaria incidence and mortality rates by at least 90% from 2016 to 2030[81]. However, control efforts have been hampered by the emergence of antimalarial drug resistance, which threatens to undo much of the progress made to-date[79,82]. As such, identifying the genomic mechanisms underlying antimalarial drug resistance is crucial.

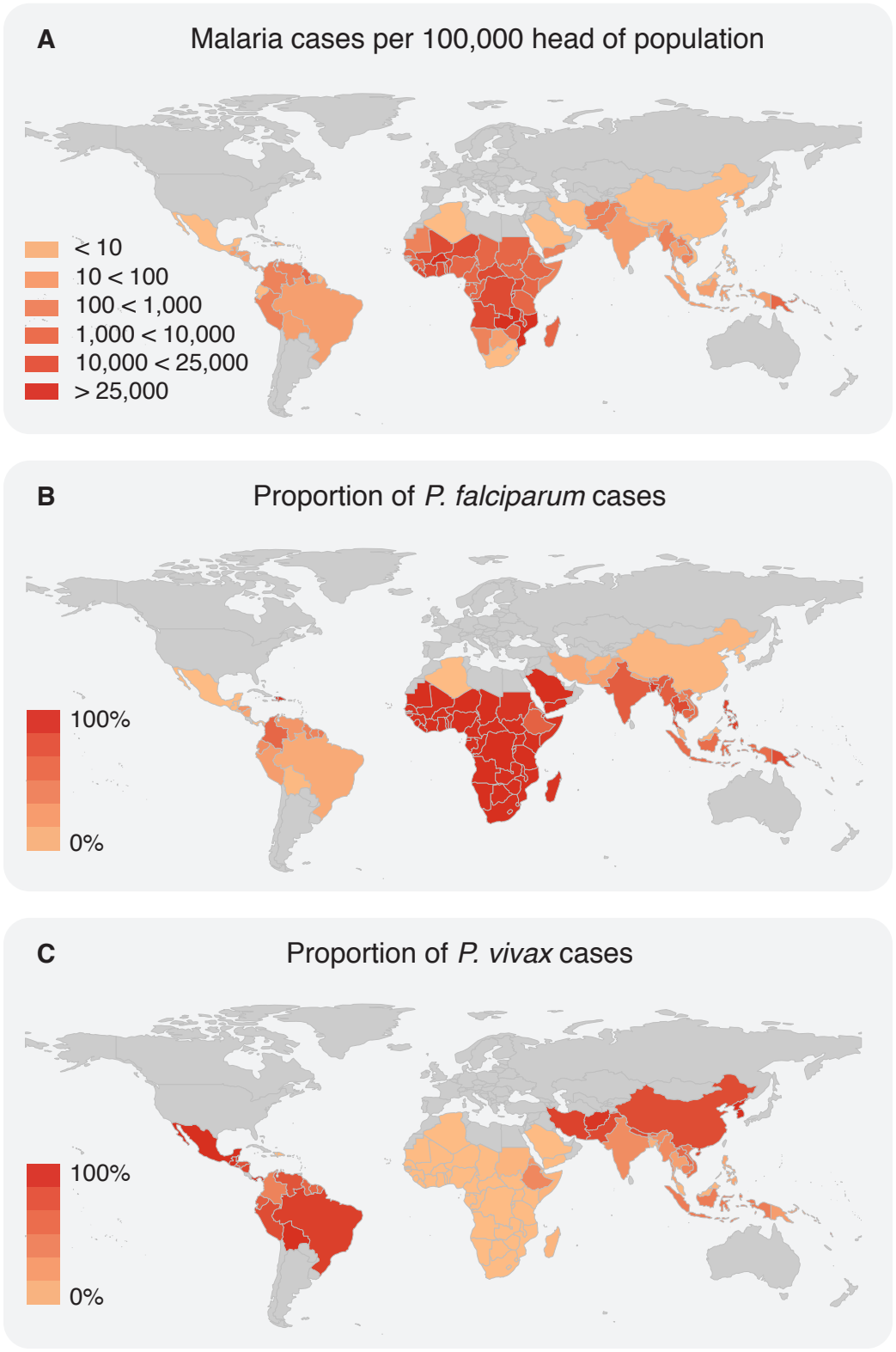


Figure 5.1: The burden of malaria across the globe. Countries colored in light-grey do not have malaria transmission. Data was sourced from http://www.who.int/malaria/publications/country-profiles/en/ for the year 2016 and plotted in R using the package rworldmap. **A** The number of reported confirmed cases of malaria, per 100,000 individuals in the population. **B** The proportion of confirmed cases of malaria that were *P.falciparum* infections. **C** The proportion of confirmed cases of malaria that were *P.vivax* infections.

In order to understand the mechanisms of resistance it is important to identify the genes that are under selection in response to antimalarial drug use. As discussed in Chapter 1, IBD analysis can be used to identify loci under positive selection and here we motivate the need for identifying loci under positive selection in the *Plasmodium* parasite. We begin by giving a brief overview of the parasite's biology and some of the challenges faced when performing genomic analyses of *Plasmodium*, followed by an introduction to antimalarial drug resistance and statistical methods for identifying selection signatures. While both *P. falciparum* and *P. vivax* are burdensome in their own right, we focus on *P. falciparum* for the remainder of this thesis.

5.1.1 The life cycle of malaria

The malaria life cycle is complicated as it involves two hosts and a number of parasite morphologies. Furthermore, the ploidy of *Plasmodium* does not remain constant. Here we explain a simplified life cycle only touching on the relevant concepts. The following was summarised from Klein[83].

Malaria is transmitted to humans through the bite of an infected female *Anopheles* mosquito when the mosquito takes a blood meal (Figure 5.2). During the blood meal, *Plasmodium* parasites are released into the human host's blood stream as haploid *sporozoites* where they make their way to the liver. Once inside the liver the sporozoites undergo asexual replication. It is during this stage of an infection where some *P. vivax* become dormant and do not asexually replicate until sometime later, possibly months or years after the initial infection resulting in a relapse infection. Following replication, the parasites burst out of the liver cells and re-enter the blood stream as *merozoites*. While in the blood stream, the merozoites rapidly invade red blood cells. Most merozoites will reproduce asexually within the cells then rupture out, destroying the red blood cell, and will continue to invade and destroy more red blood cells in this manner. This stage of the infection leads to the clinical symptoms of malaria, which include fever, headaches and anaemia. Rather than reproducing asexually within the red blood cells, a small number of merozoites will form male and female gametocytes instead; the sexual forms of the parasite.

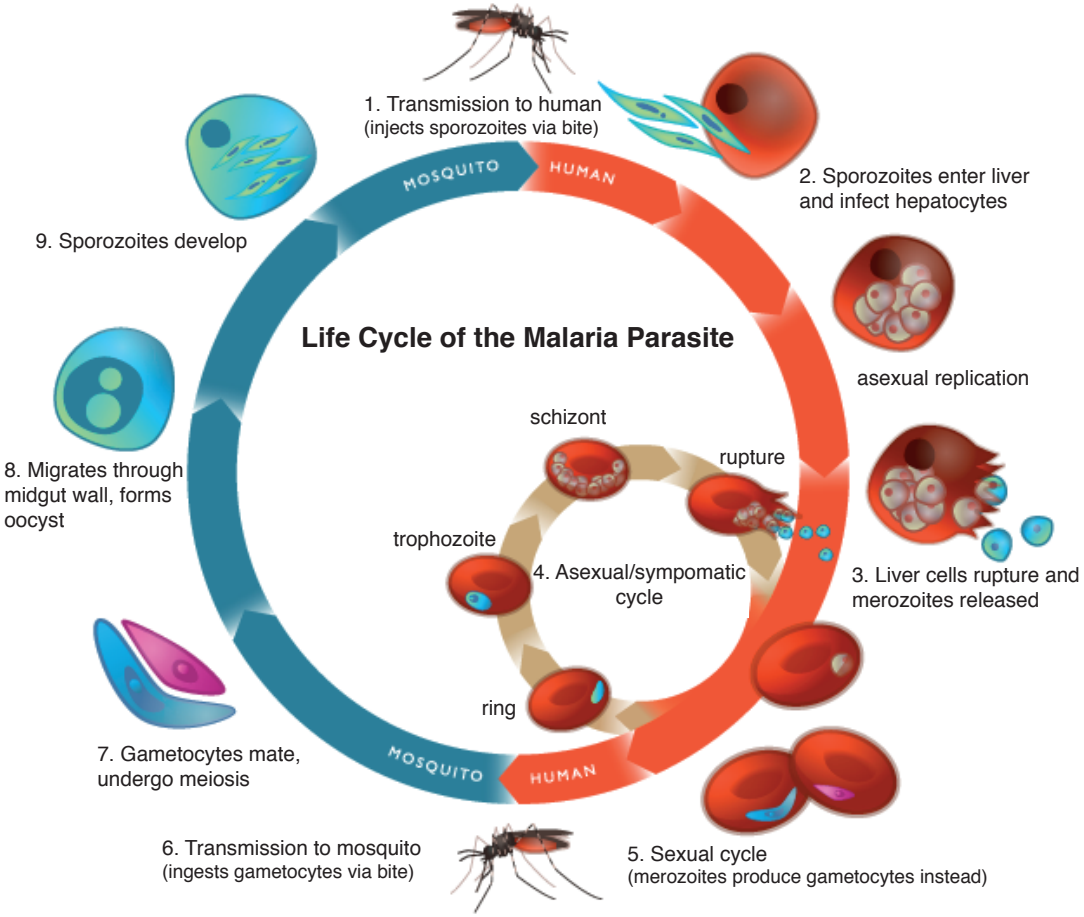


Figure 5.2: The life cycle of the malaria parasite. This image was sourced from Klein [83]. The ring, trophozoite and schizont stages were excluded from our description in the main text for simplicity.

When a mosquito takes a blood meal of a malaria infected individual, it ingests the gametocytes, which make their way to the mosquito midgut. Within the mosquito midgut the red blood cells containing the gametocytes disintegrate and the male and female gametocytes are able to fuse together to form diploid zygotes. It is here that sexual reproduction takes place, allowing for meiosis and hence recombination between the male and female gametocytes. Following meiosis, the zygotes traverse the midgut wall and develop into oocysts. Within an oocyst, sporozoites are formed, which replicate asexually. The sporozoites then burst from the oocyst and make their way to the salivary glands of the mosquito where they are ready to be released into the human host during another blood meal.

5.1.2 The *P. falciparum* genome

The genome of *P. falciparum* comprises 22.9 Mb distributed among 14 haploid* nuclear chromosomes, in addition to mitochondrial and apicoplast DNA[5]. The nuclear chromosomes range in size from 0.6 Mb to 3.3 Mb, with lengths increasing as chromosome nomenclature increases (opposite to the human genome). Unlike other species of *Plasmodium*, *P. falciparum* has an extremely high AT content with a composition of approximately 80% AT. In contrast, *P. vivax* has an AT composition more comparable to humans of 55%[5].

5.1.3 Challenges of sequencing the malaria genome

Plasmodium can be extracted from a malaria infected individual through a blood sample, where all *Plasmodium* obtained from a single blood sample constitute an *isolate*. The *Plasmodium* genome can then be sequenced using NGS technologies, however one challenge of sequencing the genome is the abundance of human DNA that is also present in the sample[84]. Such contamination can greatly reduce the coverage of the *Plasmodium* genome sequenced, resulting in poor quality data. This is further exacerbated when there is low parasitemia (parasites quantity in the blood)[84]. An alternative source of contamination is the presence of multiple species of *Plasmodium* in the isolate[85]. For example, individuals living in countries with both *P. falciparum* and *P.vivax* transmission, such as Papua New Guinea and Southeast Asia, may have both species present within an isolate, resulting from multiple infections. Contamination cannot be avoided when sequencing the *Plasmodium* genome, however pre-sequencing techniques are implemented to minimise contamination from other sources, such as separating white blood cells from red blood cells to reduce the amount of human DNA in the sample[84]. Furthermore, aligning the sequences to multiple reference genomes, including the human genome and multiple species of *Plasmodium*, can improve data quality.

Following sequence alignment and data filtering procedures, variant calling is typically performed. This may appear trivial for high coverage, good quality data, given the haploid status of the *Plasmodium* genome. However, this is not always the case as individuals can be infected with multiple, genetically-distinct strains of the same species of *Plasmodium*, giving the appearance of a ploidy > 1 isolate[86]. The number of strains contributing to an infection is termed the *multiplicity of infection* (MOI). An individual infected with a single strain has MOI = 1 while an individual infected with 3 strains has MOI = 3. Multiple

**Plasmodium* becomes diploid briefly while in the mosquito host and is otherwise haploid.

infections can arise in two ways[87];

1. Multiple mosquitos, each carrying a unique strain of *Plasmodium*, take separate blood meals of the same individual and transmit their strain.

2. A single mosquito carrying multiple strains of *Plasmodium* takes a blood meal of one individuals and transmits multiple strains.

While it is possible for an individual to be infected with many strains of the same species of *Plasmodium* in geographical regions with high malaria transmission[87], the ability to sequence all strains contributing to an infection is limited, even in infections with high parasitemia. Strains are often present at different proportions within an infection[87,88] and extracting large quantities of all strains from a single blood sample is simply not possible. However, as control efforts have intensified in recent years, the number of multiple infections has decreased, with infections commonly containing either one (MOI = 1) or two (MOI = 2) strains[87] (Figure 5.3). While MOI = 2 isolates can be treated as though they are diploid, variant callers for diploid genomes typically assume 50:50 representation of alleles in the mixture, which is not the case if MOI = 2 isolates have strains in different proportions, potentially resulting in calling errors. As such, care should be taken when processing *Plasmodium* data, and stringent filtering criterion may be required to produce a good quality dataset with confident variant calls.

In addition to the challenges of contamination and MOI, the genome of *P. falciparum* contains regions that are challenging to sequence and statistically analyse[89]. The high AT content of *P. falciparum* results in many highly-repetitive regions of the genome[5] with variable coverage and ambiguous alignments[89]. The genome also contains hypervariable gene families such as the *var*, *stevor* and *rif* genes. These are predominantly located in subtelomeric regions of the chromosomes as well as towards centromeres, and undergo ectopic recombination (recombination between non-homologous loci)[90]. As such, the genome of *P. falciparum* is highly-polymorphic at these loci and difficult to align to a reference genome. A blacklist has been created containing problematic regions of the genome, which includes highly-repetitive and highly-polymorphic regions, as well as telomeres. This constitutes approximately 10% of the genome (2.5 Mb). The remaining genome is referred to as the *core* genome[89].

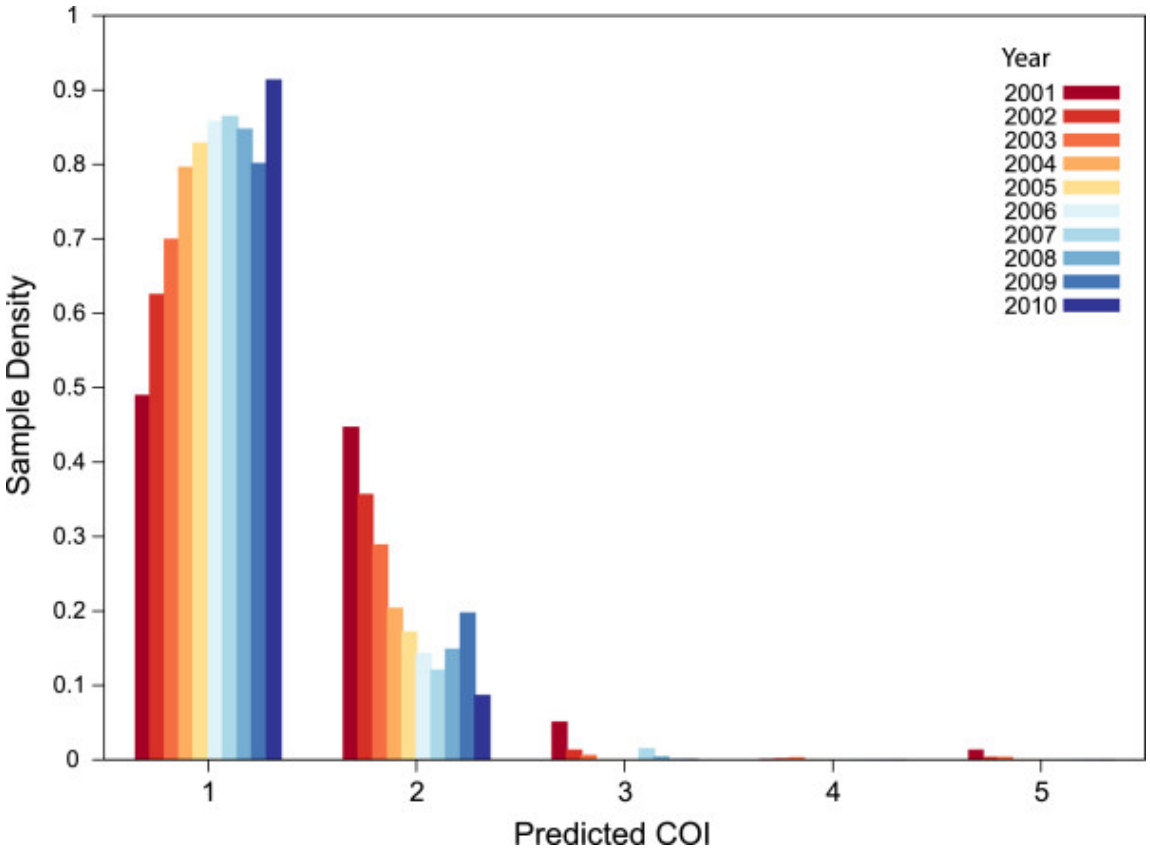


Figure 5.3: The proportion of isolates collected from Thailand between 2001 and 2010 with multiplicity of infection between 1 and 5, where MOI is denoted here as COI (complexity of infection). The total number of isolates collected was 1,731 and the image was sourced from Galinsky et al.[87].

5.2 Antimalarial drug resistance

Malaria once inhabited much of world and, with the discovery and development of a number of antimalarial drugs, has been successfully eradicated from many countries[91]. However, there is currently no vaccine for malaria and in the last 60 years' parasites have developed resistance to most antimalarial drugs used to treat infections, hampering control efforts and threatening to undo much of the progress made to date[80].

One of the most widely available antimalarial drugs, that is also considered the most successful antimalarial drug, is chloroquine. Chloroquine was introduced for treatment of malaria infections in 1946, and made considerable progress with reducing the morbidity and mortality of the disease[92]. However, within 10 years of its introduction, *P. falciparum* isolates from Cambodia were beginning to have reduced sensitivity to the drug and in 1957 chloroquine resistance was confirmed (Figure 5.4A)[92]. Specifically, the haplotype CVIET

at codons 72-76 of *Pfcrt*, the chloroquine resistance transporter gene, was found to confer with reduced sensitivity to chloroquine[93,94]. This haplotype rapidly spread throughout South and Southeast Asia, and by the 1970s, chloroquine resistance had further spread from South Asia to Sub Saharan Africa[92]. While resistance was developing throughout Southeast Asia, two independent sources of chloroquine resistance emerged in Colombia and Venezuela, carrying a second haplotype in *Pfcrt*, SVMNT. Soon after its emergence, SVMNT had swept throughout much of South America and also spontaneously emerged in Papua New Guinea[93,94]. Within 50 year of the introduction of chloroquine, resistance to the drug was present in almost every country with malaria transmission, and the efficacy of chloroquine as an antimalarial drug reduced considerably.

In response to the emergences of chloroquine resistance, new antimalarial drugs were developed, including sulphadoxine and pyimethamine, mefloquine and piperaquine. However, it was not long before resistance to these drugs also emerged[80].

More recently artemisinin was introduced as an antimalarial drug and has been used in combination with partner drugs, including sulphadoxine and pyimethamine, mefloquine and piperaquine, for treatment of *P. falciparum* infections[80]. Artemisinin combination therapies (ACT) are recommended as the first-line treatment for malaria infections in countries with endemic malaria and have been crucial in the recent developments with reducing the global burden of this disease[79]. Artemisinin was first introduced in 2002, however in 2007 resistance was reported in Cambodia[95] and has since emerged in 5 countries in the greater Mekong subregion[96] (Figure 5.4B). More than 20 point mutations in *Pfk13*, the kelch13 propeller domain, have been associated with artemisinin resistance on a number of different haplotype backgrounds[97]. The emergence of artemisinin resistance has been described as a global health crisis[98] and there are growing concerns of resistance developing in Africa, where more than 90% of malaria deaths occur. Furthermore, it is feared that *Plasmodium* may develop resistance to multiple drugs simultaneously, in which case the partner drug used in combination with artemisinin would be inefficient.

Programs have been developed in an attempt to monitor and control the spread of artemisinin resistance[80], however, if unsuccessful, could have catastrophic consequences for the progress of malaria elimination efforts.

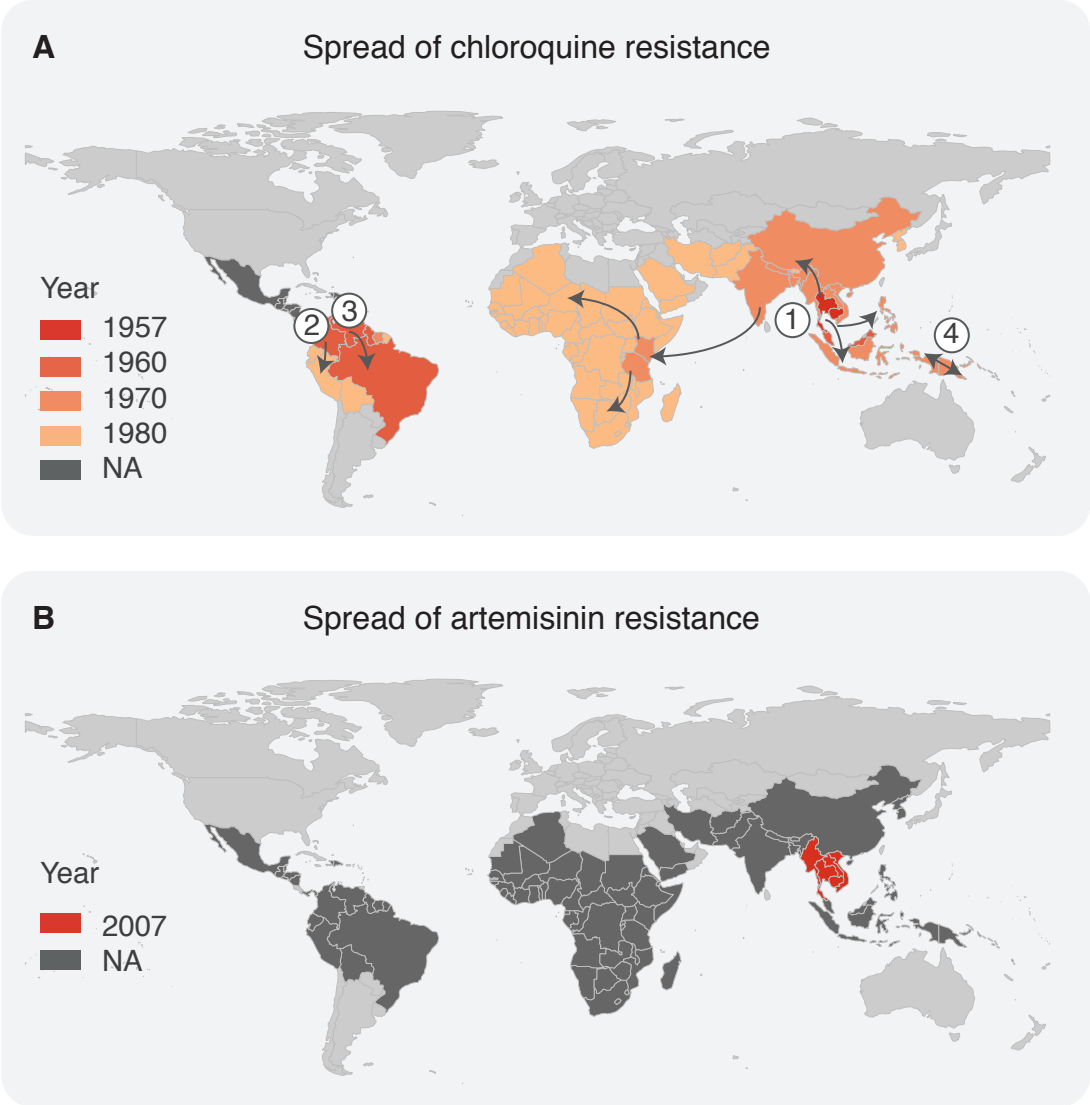


Figure 5.4: A timeline of the emergence and spread of antimalarial resistance. **A** The emergence and spread of chloroquine resistance. Data was sourced from[98]. **B** The emergence and spread of artemisinin resistance. Data was sourced from WHO[99].

5.3 Selection of antimalarial drug resistant variants

Chloroquine resistance emerged independently in four geographical locations on two haplotype grounds[93,94]. The short time interval over which resistance to chloroquine spread and the increase in resistance-haplotype frequencies is consistent with a hard-selective sweep. In contrast, the emergence of a great number of artemisinin-resistant variants on many haplotype backgrounds at low frequencies is consistent with a soft selective sweep[97]. In order to advance malaria control and elimination, it is crucial to identify loci like *Pfcrt* and *Pfk13* that are under positive selection and are associated with antimalarial drug

resistance.

5.3.1 Methods for identifying selection

There are a number of methods that can identify loci under selection that differ in the nature of the selection signature they aim to identify, which in turn depends on the time since the selection pressure began[100]. When an allele is selected for, it increases in frequency in the population, along with neighbouring alleles, such that the stretch of genome surrounding the selected allele is relatively homogeneous. As more time passes and the allele reaches high frequencies, new alleles neighbouring the favoured allele are introduced, initially at low frequencies as they have only recently appeared[101]. This type of positive selection can be readily detected by allele frequency-based methods like Tajima's D[102], that operate by identifying genomic regions with an abundance of rare alleles, consistent with more ancient positive selection[100]. Tajima's D is ideal for detecting hard sweeps where the selected allele's frequency is near fixation, however does not perform well when selection acts on standing variation or in the case of a soft-selective sweep[21,103].

Alternatives to frequency-based methods are LD-based methods. These methods operate by identifying haplotypes that are unusually long relative to their frequency in the population[101]. The premise here is that recently selected alleles are situated on long haplotypes as there have not been many recombination events to shorten the haplotypes and break down LD. A number of LD-based methods initially utilize the extended haplotype homozygosity (EHH) method, which simply measure the amount of LD-decay around a core haplotype[104]. As the distance from the core haplotype increases, EHH decreases. A core haplotype with an unusually high EHH and population frequency is indicative of recent positive selection. A popular method that uses EHH is the integrated haplotype score (iHS)[101]. iHS compares the decay of EHH between two alleles (the derived allele and the ancestral allele) at a given locus, by calculating the area under the curve defined by EHH for each allele. A region with slowly decaying EHH for the derived allele, relative to the ancestral allele, provides evidence for recent positive selection. Methods that make use of EHH can readily identify a hard selective sweep, however have reduced power to identifying selection on standing variation or as a soft selective sweep[21,105]. This is because these types of selection are not the result of a single, long haplotype that is common in the population. Rather, multiple haplotype backgrounds are present as a result of novel variants increasing haplotype diversity in addition to recurrent variants[21].

An alternative LD-based method is IBD analysis, which identifies loci under positive selection if an abundance of IBD is detected, relative to other loci in the genome[21]. Like other LD-based methods, IBD analysis is ideal for hard selective sweeps, however it also has high power to detect selection acting on standing variation and as a soft selective sweep[21,105]. This is because IBD analysis can take advantage of IBD segments inferred between different pairs of individuals, reflecting genetically diverse haplotypes, although genetic signatures may not be as prominent as with hard selective sweeps.

5.3.2 Detecting selection in malaria

Antimalarial drug resistance is a relatively recent occurrence. Therefore, methods for identifying loci under selection in *Plasmodium* should be intended for recent selective pressures. Additionally, methods should be able to determine selection acting on standing variation as well as hard and soft selective sweeps, as these types of selection have been found to play a role in antimicrobial resistance[97,106,107]. Given this criterion, LD-based methods, such as IBD analysis and iHS, appear to be the most appropriate methods for this task.

Although IBD analysis has higher power to detect more complicated selective sweeps than iHS[21,105], selection signatures are most commonly identified in malaria using the iHS methodology[108,109,110,111]. This is unfortunate as iHS requires phased haplotype data to determine selection, which is problematic for isolates with MOI > 1, as information on the number of strains in an infection and the respective proportions that each strain contributes to the infection must be known. This information is not trivial to extract from sequencing data and as such, isolates with MOI > 1 are typically excluded from analysis, which can greatly reduce the power of an analysis[86].

Part of the reasoning behind the regular use of iHS as opposed to IBD analysis for positive selection in *Plasmodium* is the lack of IBD methodologies for haploid species. In fact, there have been no reported IBD analyses performed on haploid microorganisms to identify loci under positive selection. Furthermore, none of the available selection methods are equipped to handle the added complexities that arise from multiple infections. As such, the development of an IBD tool that is specifically designed for haploid species that can accommodate multiple infections would be beneficial for disease elimination and control efforts of malaria, and other diseases.

In the next chapter I describe how the framework for IBD detection developed in

Chapter 2 can be modified for plasmodium IBD detection in the presence of MOI ≥ 1, and describe a new selection statistic that makes use of IBD signals.

Chapter 6

Detecting selection signals in *P. falciparum* using IBD analysis

This chapter has been submitted for publication and is currently under review. The manuscript is publicly available on BioRxiv[112] and has been re-formatted to meet the requirements of this thesis.

6.0.1 Background

The progress of malaria control and elimination efforts is under threat with the emergence of antimalarial drug resistance[79]. As discussed in Chapter 5, there is a need for a methodology that can readily identify positively-selected loci, in the form of multiple sweeps, that have arisen due to antimalarial drug pressure. Furthermore, a methodology that is suitable for isolates with multiple infections is desirable to avoid the reduction in power that results from excluding such isolates from analysis.

In Chapter 6 we introduce isoRelate, a freely available R package (https://github.com-bahlolab/isoRelate) that performs IBD analysis on recombining haploid species, such as the malaria-causing parasite *Plasmodium* and the bacterium *Staphylococcus aureus*, that also includes multiple infections. Unlike other selection methods such as iHS[101], IBD mapping of microorganisms can also be used to infer fine-scale population structure and allows the ability to monitor disease control and transmission, as well as to determine if an antimicrobial drug-resistant haplotype has spread or arisen independently at different geographical locations. Furthermore, IBD mapping has the potential to uncover multidrug resistance and, for diseases that experience relapse infections such as malaria caused by

P. vivax, may be able to distinguish between new or relapsing infections in drug efficacy and cohort studies.

Using isoRelate, we demonstrate the ability of IBD analyses to detect signals of recent positive selection using WGS data for a previously published global *P. falciparum* dataset of 2,550 isolates[97]. We make comparisons with other popular methodologies that also try to detect recent positive selection. Additionally, we use isoRelate to explore *P. falciparum* population structure between geographical regions; confirm the global spread of resistance to the antimalarial drug chloroquine as well as explore resistance to artemisinin as a soft selective sweep, and investigate the ability of IBD to detect multidrug resistance.

6.1 Datasets

6.1.1 MalariaGEN genetic crosses dataset

To validate our method's ability to recapitulate recombination events and thus IBD sharing we made use of a previously published *P. falciparum* genetic cross. WGS data was retrieved for 98 *P. falciparum* lab isolates that were generated as part of the MalariaGEN consortium Pf3k project[89]. This dataset included the parent and progeny (first generation) of crosses between the pairs of parent strains 3D7 and HB3, 7G8 and GB4, and HB3 and Dd2. We retrieved all available Pf3k data in VCF file format from data release 5 (https://www.malariagen.net/data/pf3k-5). SNPs were excluded if they were not in a core region of the genome[89], or if they had Quality of Depth ≤ 15 or Mapping Quality ≤ 50, or if less than 90% of samples were not covered by at least 5 reads, or they were not polymorphic or if their MAF was less than 1% (using a read depth estimator). Samples were also excluded if less than 90% of their SNPs were not covered by at least 5 reads. Appendix C Table 1 shows the number of isolates and SNPs before and after filtering of each genetic cross.

We visualized parental recombination breakpoints in the progeny using the haplotypes displayed in the online data application (https://www.malariagen.net/apps/pf-crosses/1.0/). We selected haplotypes that were constructed following GATK variant calling[113] with all other in the online application parameters at default values[89]. This allowed us to produce a gold standard IBD datasets with known recombination events. We then assessed isoRelate's inferred IBD segment locations against this dataset.

6.1.2 MalariaGEN global *P. falciparum* dataset

WGS was performed on 2,512 *P. falciparum* field isolates sampled from 14 countries across Africa and Southeast Asia as part of the MalariaGEN consortium Pf3k project[97,114]. We retrieved all available Pf3k data in VCF file format from release 5. We merged all nuclear chromosome VCF files and applied filters to the 2,512 samples and 1,057,870 biallelic SNPs.

Variants were filtered using GATK's SelectVariants and VariantFiltration modules[113]. SNPs were excluded if there were more than 3 SNPs within a 30 base pair window, or if they were not in a core region of the genome, or if they had Variant Quality Score Recalibration (VQSR) < 0. Moreover, to reduce the possibility of spurious SNP calls further filters for Quality of Depth (QD), Strand Odds Ratio (SOR), Mapping Quality (MQ) and MQ Rank Sum (MQRankSum) were applied (QD > 15, SOR < 1, MQ > 50, MQRankSum > -2). This filtering left 561,695 SNPs in the dataset.

Next, separating the data by country of origin, SNPs were excluded if less than 90% of samples were not covered by at least 5 reads or they were not polymorphic. Samples were also excluded if less than 90% of their SNPs were not covered by at least 5 reads. Following this, countries were grouped into broader geographical regions of West Africa, Central Africa or Southeast Asia, and the intersection of SNPs within a region was taken. Lastly, within each country, SNPs with MAF less than 1% (using read depths) were removed. Appendix C Table 2 displays the number of isolates and SNPs before and after filtering of each country. Nigeria was excluded from all downstream analyses due to the low number of SNPs remaining after filtering.

6.1.3 Papua New Guinea dataset

WGS data was available for 38 *P. falciparum* isolates from Madang, Papua New Guinea, sampled in 2007 and sequenced at the Wellcome Trust Sanger Institute (WTSI), Hinxton, UK as part of the MalariaGEN consortium (http: //www.malariagen.net/about; study ID: 1021-PF-PG-MUELLER). The sequencing data was processed by replicating the analysis processing steps of the MalariaGen Pf3k field isolates for compatibility (Appendix C).

6.1.4 Simulated data with known selective sweeps

To assess the ability of IBD to detect the selective sweeps illustrated in Figure 1.5, we simulated SNP data in the presence of various sweeps using the forward population genetic

simulator, SLiM[115], under an evolutionary model for *P. falciparum*. Specifically, we simulated a 2.27 Mb region, which is approximately the length of *P. falciparum* chromosome 12, under four different scenarios; no selection and positive selection via hard sweeps, soft sweeps and standing variation.

We generated an initial population that resembles *P. falciparum* assuming a constant effective population size of 100,000[116], a mutation rate of 1.7×10^{-9} per base pair per generation[117] and a recombination rate of 7.4×10^{-7} per base pair per generation[89]. The forward simulation was run over 400,000 generations, after which a sample of 10,000 haplotypes was randomly drawn to undergo selective pressures as follows. We note that it would have been desirable to run the simulation over more generations[116], however this was not computationally feasible with the forward simulator.

Under the scenario of no selection, SLiM was run on the sampled population with all alleles having the same fitness (i.e. selection coefficient s = 0). A hard sweep was generated by sampling one haplotype to introduce a new allele with a selection coefficient of either $s = 0.01$, $s = 0.1$ or $s = 0.5$. Alternatively, selection on standing variation was introduced by adding a selective advantage of $s = 0.01$, $s = 0.1$ or $s = 0.5$ to an existing allele with a population frequency of either $f = 0.01$, $f = 0.05$ or $f = 0.1$. Finally, soft sweeps were generated such that a new allele would arise and spread throughout the population on multiple haplotype backgrounds. We introduced the new allele at random generations, where, at each generation, one haplotype was sampled that was not already carrying the allele, and the allele was inserted. For each soft sweep, the selected allele had identical selection coefficients on each haplotype of either $s = 0.01$, $s = 0.1$ or $s = 0.5$. The number of generations between the introduction of the new allele was randomly sampled from a Poisson distribution with mean 3 generations. The allele was introduced a total of 30, 10 and 5 times over the course of each soft sweep for selection coefficients $s = 0.01$, $s = 0.1$ and $s = 0.5$, respectively. We needed to introduce the allele on more haplotype backgrounds when smaller selection coefficients were used as we wanted multiple haplotypes to sweep through the population without the allele being lost straight away. We generated 10 replicates for each scenario (no selection = 1, hard sweep = 3, standing variation = 9, soft sweep = 3), randomly assigning the genetic position of the selected allele, and sampled 200 haplotypes at generations 50, 100, 200 and 500 following the initial sampling of the population, resulting in a total of 150 simulated datasets. The dominance coefficient of all selective sweeps was 1.

6.2 Methods

6.2.1 Assessing MOI

We applied the F_{ws} metric, a characterization of within host diversity, to each countries SNP sets to determine isolates that had multiple infections[114]. An isolate was classified as having multiple infections if $F_{ws} < 0.95$. For each country PED and MAP files for downstream analysis were extracted using moimix[118]. Heterozygous SNP calls were retained for isolates assigned as having MOI greater than 1, otherwise heterozygous SNPs were set to having a missing value at those SNPs to signify the likelihood of a genotyping error.

6.2.2 IBD detection and segment filtering

The methodology implemented in isoRelate is identical to that described in Chapter 2 for XIBD model 1, and here we detail the specifications required for analysis of isolates with MOI ≥ 1. The extension to haploid species with multiple infections essentially concerns replacing the sex of the individuals in XIBD with MOI status. An isolate with MOI = 1 consists of a single strain and is analyzed as if it were haploid; thus sharing either 0 or 1 allele IBD with any other isolate. An isolate with MOI > 1 consists of multiple genetically distinct (and possibly related) strains, and is considered diploid; sharing 0, 1 or at most 2 alleles IBD with other isolates. Here we make the assumption that an isolate with MOI > 1 actually has MOI = 2, arguing that the current coverage of WGS data struggles to identify more than two clones contributing to an isolate. This assumption will be incorrect for some isolates; however, the progress of malaria control efforts has lead to a decrease in the number of multiple infections, with the majority of multiple infections consisting of two strains[87] (Figure 5.3).

We compute the allele frequencies for each country separately for *P. falciparum*. This is necessary due to the highly divergent sets of SNPs observed in *P. falciparum* globally[119]. To perform IBD analyses between isolate from different countries, SNPs were included in the analysis if the population allele frequencies between the pair of countries differed by less than 0.3. A MAF concordance threshold of 0.3 was arbitrarily used in the analysis as this threshold resulted in the inclusion of at least 75% of SNPs present in both populations, for all pairwise-population comparisons. Population allele frequencies for the combined countries were then calculated using all isolates from pairs of countries being examined. SNPs with MAF less than 1% were removed from the analysis along with SNPs with

missing genotype data for more than 10% of isolates. Similarly, isolates with missing genotype data for more than 10% of SNPs were removed and a genotyping error rate of 1% was included in the model. Appendix C Tables 2 and 3 give the number of isolates and SNPs before and after filtering for each country and pairwise-country dataset.

IBD segments that contain less than 20 SNPs or have lengths less than 50,000bp are excluded, as they are likely to represent distant population sharing that is not relevant to recent selection. IBD analyses were performed between all pairs of isolates that remained once filtering procedures had been applied.

6.2.3 Identifying selection signals and assessing significance from IBD

In order to assess cohort level IBD sharing and thus investigate selection, we developed a test statistic that overcomes some of the limitations of other selection statistics. In particular, we developed a statistic that does not require phased data and that takes into account relatedness observed between isolates. This test statistic is better suited for analyses of microorganisms, like Plasmodium, that experience strong positive selection of various types in addition to mixed infections. Using ideas previously applied in algorithms such as EIGENSTRAT [120] we were able to derive a test statistic that showed approximate normality and thus can be interpreted probabilistically using distributional assumptions. The test statistic was calculated as follows:

We created a matrix of binary IBD status with rows corresponding to SNPs and columns corresponding to isolate pairs. For each column, we subtract the column mean from all rows to account for the amount of relatedness between each pair. Following this we subtract the row mean from each row and divide by the square root of $p_i(1-p_i)$, where p_i is the population allele frequency of SNP i. This adjusts for differences in SNP allele frequencies, which can affect the ability to detect IBD. Next we calculate row sums and divide these values by the square root of the number of pairs. These summary statistics are then normalized genome-wide such that they follow a standard normal distribution with a mean of 0 and standard deviation of 1. Negative z-scores are difficult to interpret when investigating positive selection; therefore we square the z-scores such that the new summary statistics follow a chi-squared distribution with 1 degree of freedom (Figure 6.1). This produces a set of genome wide test statistics $\{X_{iR,s}\}$, where $X_{iR,s}$ is the chi-square distributed test statistic for IBD sharing from isoRelate at SNP s.

We calculate p-values for $\{X_{iR,s}\}$, after which we perform a $-\log 10$ transformation of

the p-values to produce our final summary statistics, used to investigate the significance of selection signatures. Finally, a 5% genome-wide significance threshold was used to assess evidence of positive selection.

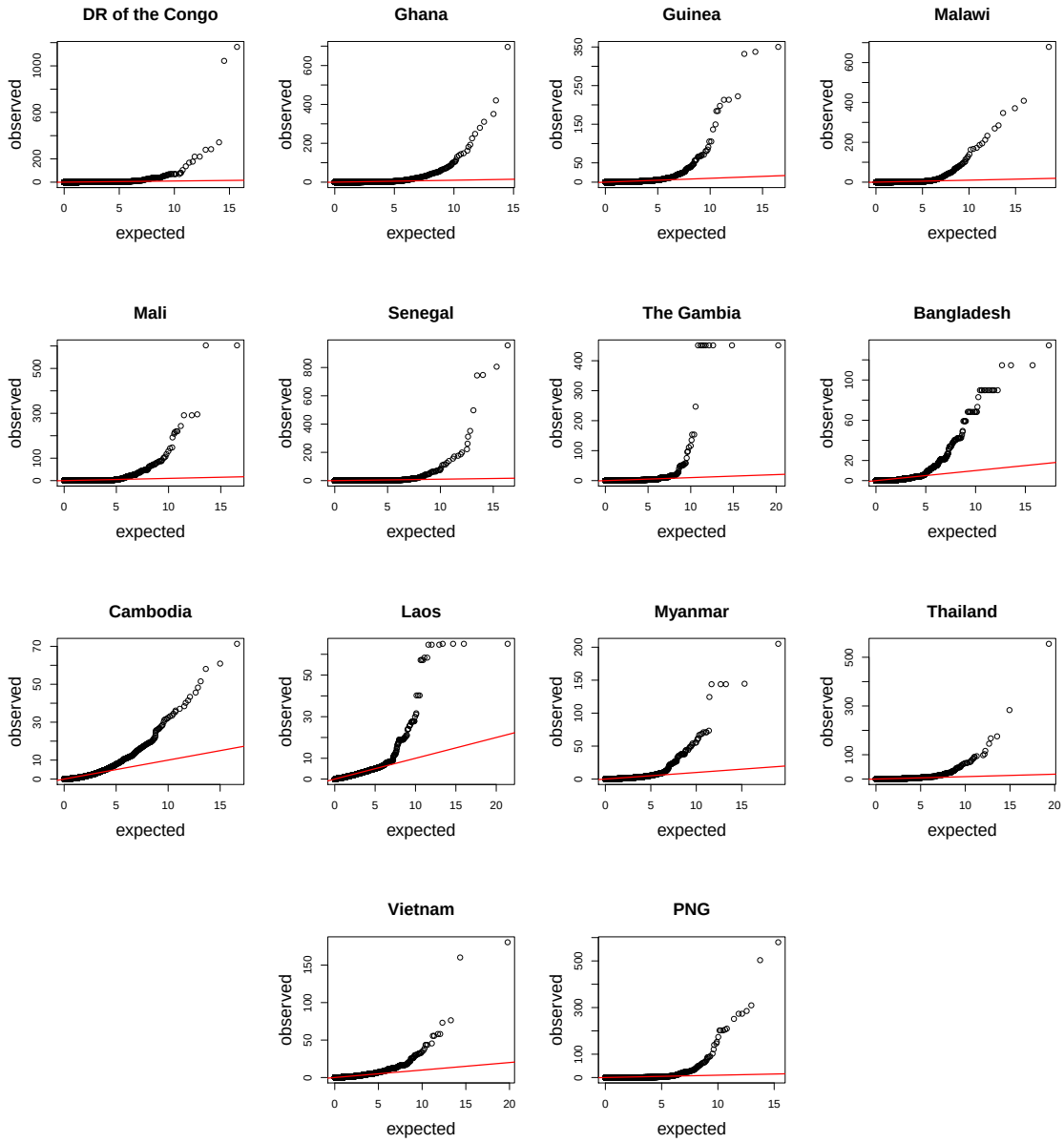


Figure 6.1: Chi-square quantile-quantile plots for the normalization step in the calculation of $\{X_{iR,s}\}$ for the global *P. falciparum* dataset. Expected quantiles ($\chi^2(df = 1)$) are on the x-axis and sample quantiles are on the y-axis.

6.2.4 Comparing methods for the detection of selection

We performed a standard analysis of selection signals using the scikit-allel v0.201.1 package in Python 2.7[121,122]. To compute selection statistics on simulated data we calculated the

iHS for SNPs passing a MAF filter of 1%[101]. We note that SNPs were removed from analysis if they were not in a core region of the genome as defined by Miles et al.[89]. We report the iHS if the EHH[104] decays to 0.05 before reaching the final SNP examined within a maximum gap distance of 2 Mb spanning the EHH region, otherwise iHS was set to missing. To standardize iHS we binned all SNPs into 100 equally sized bins partitioned on allele frequencies and then subtracted the mean and divided by the standard deviation of iHS within that bin. We computed $\log 10$ p-values using the normalized iHS from a standard normal distribution.

To detect selection using haploPS[123], SNPs passing a MAF filter of 1% that were in core regions of the genome were analysed. We first calculated the adjusted haploPS score for haplotypes identified at core frequencies of 5% to 95% in increments of 5%. This score is calculated by comparing the lengths of the identified haplotypes to the lengths of other haplotypes that are present as similar frequencies in the dataset. Regions were considered to be under positive selection if the adjusted haplotype score was less than 0.05. Since haplotypes are identified across multiple core frequencies, similar regions of positive selection are detected across these frequencies. We stacked the significant haplotypes around each SNP, identified across the different core frequencies, and calculated the number of significant haplotypes that overlap each SNP. Regions that have undergone strong positive selection in the form of a hard sweep will typically be inferred as positively selected across multiple core frequencies, therefore the number of significant haplotypes that overlap each SNP within these regions should be larger than those in regions that have not undergone selection.

Since a large number of analyses were carried out (10 replications for each of the 15 scenarios of sweeps, with haplotypes sampled at 4 time points following selection), results were summarised as follows. For isoRelate and iHS, we calculated the genetic distance between the SNP with the largest $-\log_{10}$ p-value and the selected allele. While for haploPS we calculated the distance between the selected allele and the SNP with the most number of significant haplotypes inferred across the core frequencies. Boxplots were created for each combination of scenarios from the 10 replications. Boxplots centered around zero with a small interquartile range are indicative of a sweep being consistently detected, and a method performing well.

6.2.5 Relatedness networks

To examine the haplotype sharing between isolates within and between countries, both as genome-wide averages and at a regional level, we generated relatedness networks using the R package igraph[124]. Each node in the network represents a unique isolate and an edge is drawn between two nodes if the isolates are IBD anywhere within interval. Isolates with MOI = 1 are represented by circle nodes while isolates with MOI > 1 are represented by squares. Node colors are unique for isolates from different countries.

6.2.6 Detecting multidrug resistance

To investigate multidrug resistance, whereby parasites are resistance to multiple antimalarial drugs, we extract all pairs who are IBD over a drug resistant gene of interest, $gene_1$. Here a pair is classified as IBD if they have an IBD segment that partially or completely overlaps $gene_1$. From this subset of pairs, we calculate our selection signal, X_{iR}, as per usual and investigate the distribution of these statistics across the genome. We examine all loci with significant X_{iR} for known antimalarial drug resistant genes. If resistant genes are identified in any of the loci, then we take this as evidence of joint-inheritance of these genes with $gene_1$. This subset of pairs can then be examined for resistance haplotypes in $gene_1$ and associated genes for evidence of multidrug resistance.

6.3 Results

6.3.1 Validation of isoRelate

We validated our methodology by applying isoRelate to the MalariaGEN Pf3k genetic cross dataset[89] to detect known recombination events. This dataset contains the parents and offspring of three P. falciparum strain crosses; 3D7 x HB3, 7G8 x GB4, and HB3 x Dd2. There are 21, 40 and 37 isolates for the three crosses respectively, and 11,612 SNPs, 10,903 SNPs and 10,637 informative SNPs remaining following filtering procedures (Appendix C Table 1). We combined the results for all three crosses and found that isoRelate detected 98% of all reported IBD segments, with an average concordance between inferred and reported segments of 99%. Additionally, isoRelate detected segments with 99% accuracy; meaning only 1% of segments were likely to be false positives. We did not infer IBD between any of the founders. This is expected given the documented origins of these three strains, which were derived from very different geographic regions[114]. False negatives, where IBD was not inferred between parents and offspring, were observed predominantly in genomic regions located between recombination events. Moreover, identical segment boundaries were detected between all replicate isolates. We note that our methodology has been extensively tested on simulated data for the human X chromosome and as such we have not performed simulation studies here[63].

6.3.2 Analysis of selection signal methodologies on simulated data

Multiplicity of infection = 1

We compared the selection signatures generated by isoRelate to those detected by the integrated haplotype score (iHS)[101] and haploPS[123], where iHS makes use of the EHH[104] and is designed to identify strong signals of recent positive selection, while haploPS determines strong positive selection by comparing the lengths of identified haplotypes with other haplotypes genome-wide at similar frequencies. Both iHS and haploPS require knowledge of haplotype phase, which is currently not possible for isolates with MOI > 1 as deconvolution of the contributing haplotypes for each isolate is currently not feasible. In contrast, isolates with MOI $= 1$ are derived from single haploid haplotypes. Therefore we performed initial comparisons of isoRelate, iHS and haploPS using only isolates with MOI $=$

1. isoRelate and iHS produce selection statistics that follow known distributions. We thus generated quantile-quantile plots for SNP specific test statistics for both of these methods (Appendix C Figures 1 - 2).

No method is able to detect a sweep with a selection coefficient of $s = 0.01$, regardless of the type of sweep (Figures 6.2 - 6.6). Alternatively, sweeps with selection coefficients of $s = 0.1$ and $s = 0.5$ are more readily identified. While it is unfortunate that sweeps with small selection coefficients are not detected, we anticipate that selection coefficients for variants associated with antimalarial drug resistance will be larger than 0.01 as parasites carrying the resistant variants are far more likely to survive drug treatment and reproduce.

For analysis of hard sweeps, HaploPS outperforms isoRelate and iHS, particularly as the selection coefficient increases (Figure 6.2). Specifically, haploPS is able to detect a hard sweep with selection coefficient s 0.1 at least 500 generations after its introduction while isoRelate and iHS are limited to less than 500 generations. In contrast, isoRelate and iHS are better able to detect a soft selective sweep than haploPS, with comparable performances to a hard sweep (Figure 6.3). This is surprising as iHS it is expected to have reduced performance for sweeps on multiple haplotype backgrounds[21,34]. However, data was simulated such that the selected allele (i.e. the derived allele) is identical on all haplotype backgrounds in a soft sweep (i.e. recurrent mutations). This means that iHS should still detect a soft sweep of this kind as the decay of EHH is compared between the derived allele and the ancestral allele. Only one haplotype carrying the derived allele is examined when iHS is calculated, therefore the results should be similar to those of a hard sweep with the same selection coefficient.

As expected, selection on standing variation is better detected when the initial frequency of the selected allele is low (Figures 6.4 - 6.6). Nonetheless, haploPS has limited ability to detect selection on standing variation, even with an initial allele frequency of 1%. In contrast, isoRelate has the greatest ability to detect selection on standing variation, although this is limited to less than 200 generations after the sweep is introduced.

Across all scenarios of positive selection considered here, isoRelate has the greatest ability to detect a sweep that occurred less than 200 generations after its introduction. Hughes and Verra (2001) used three generations per year as a conservative estimate of the average generation time in P. falciparum. Given this, isoRelate should be able to detect sweeps that occurred up to approximately 66 years ago, depending on selection coefficient, which is within the timeframe of reported antimalarial drug resistance.

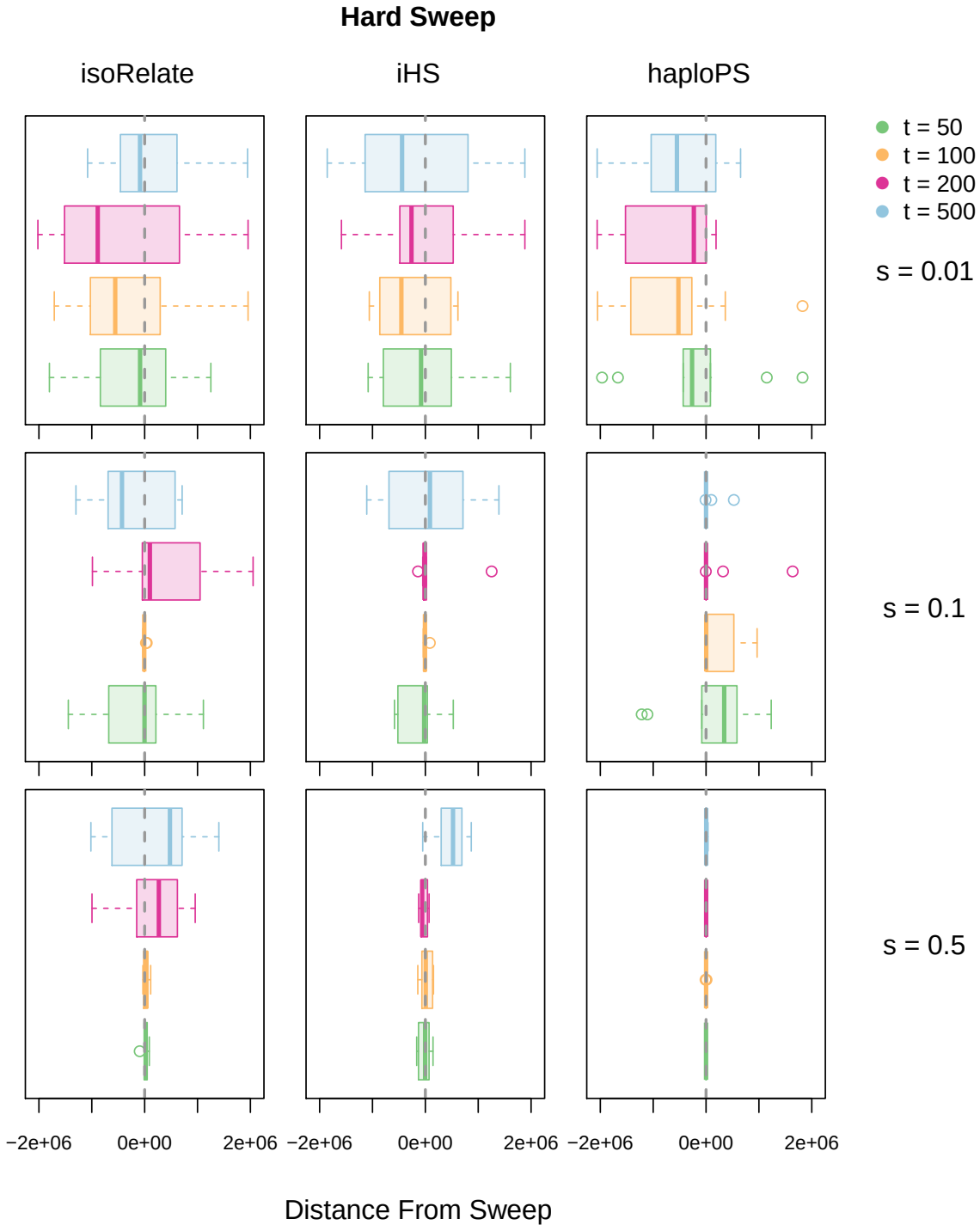


Figure 6.2: Simulation results from hard sweeps for different selection coefficients. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest $-\log_{10}$ p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario. Boxplots centered around zero with a small interquartile range indicate sweeps that are consistently detected.

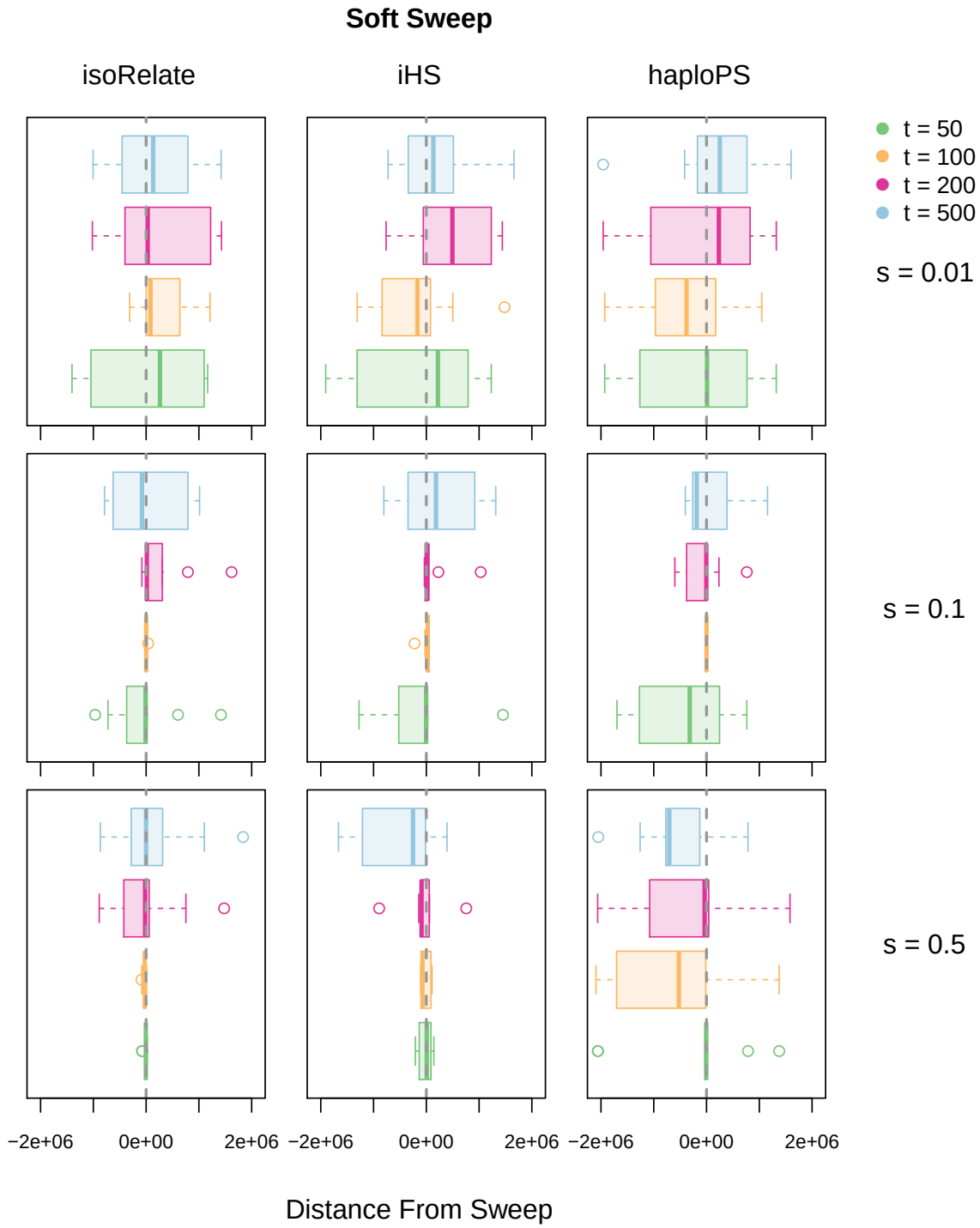


Figure 6.3: Simulation results from soft sweeps for different selection coefficients. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest $-\log_{10}$ p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario.

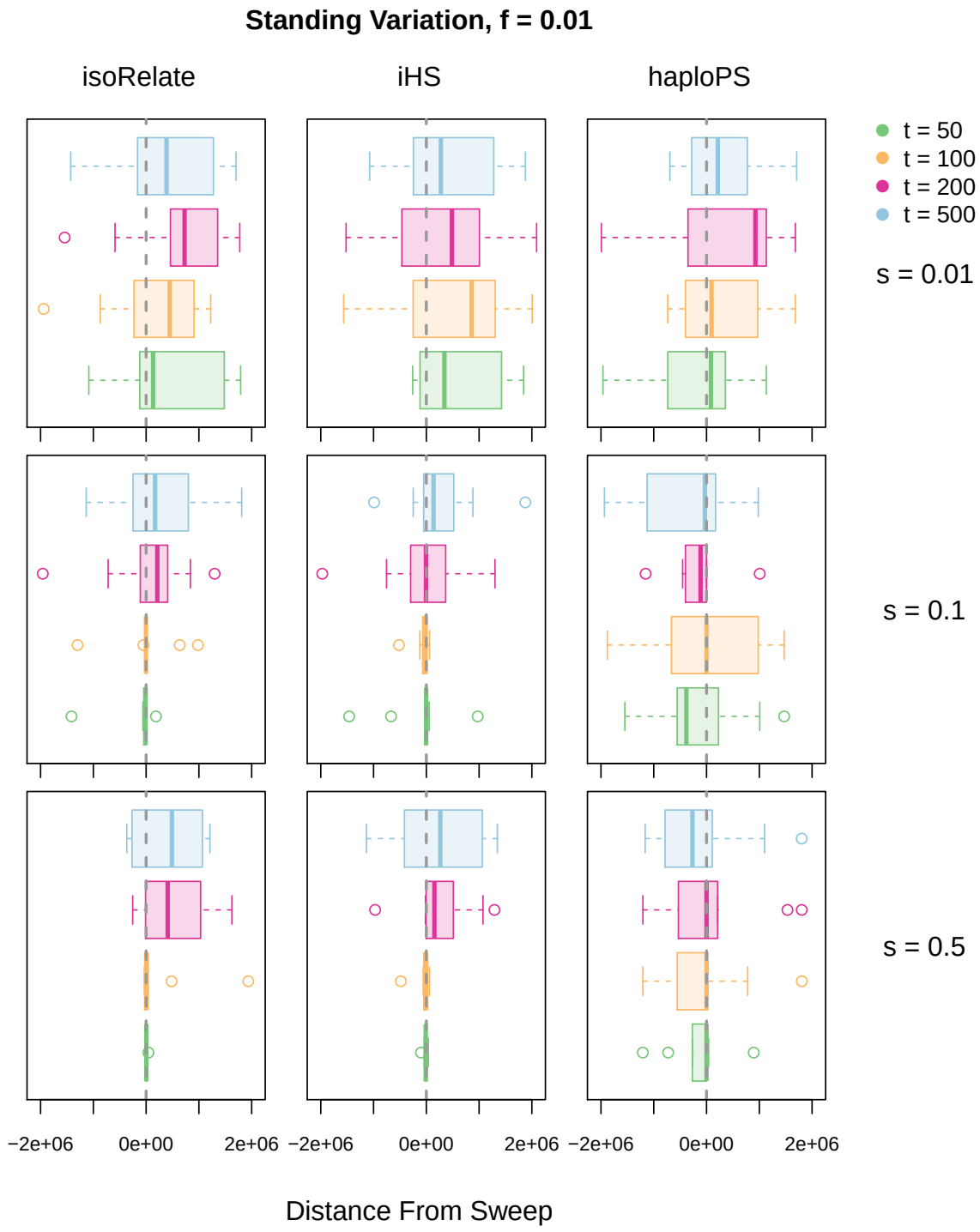


Figure 6.4: Simulation results from standing variation with initial allele frequency $f = 0.01$ for different selection coefficients. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest $-\log_{10}$ p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario.

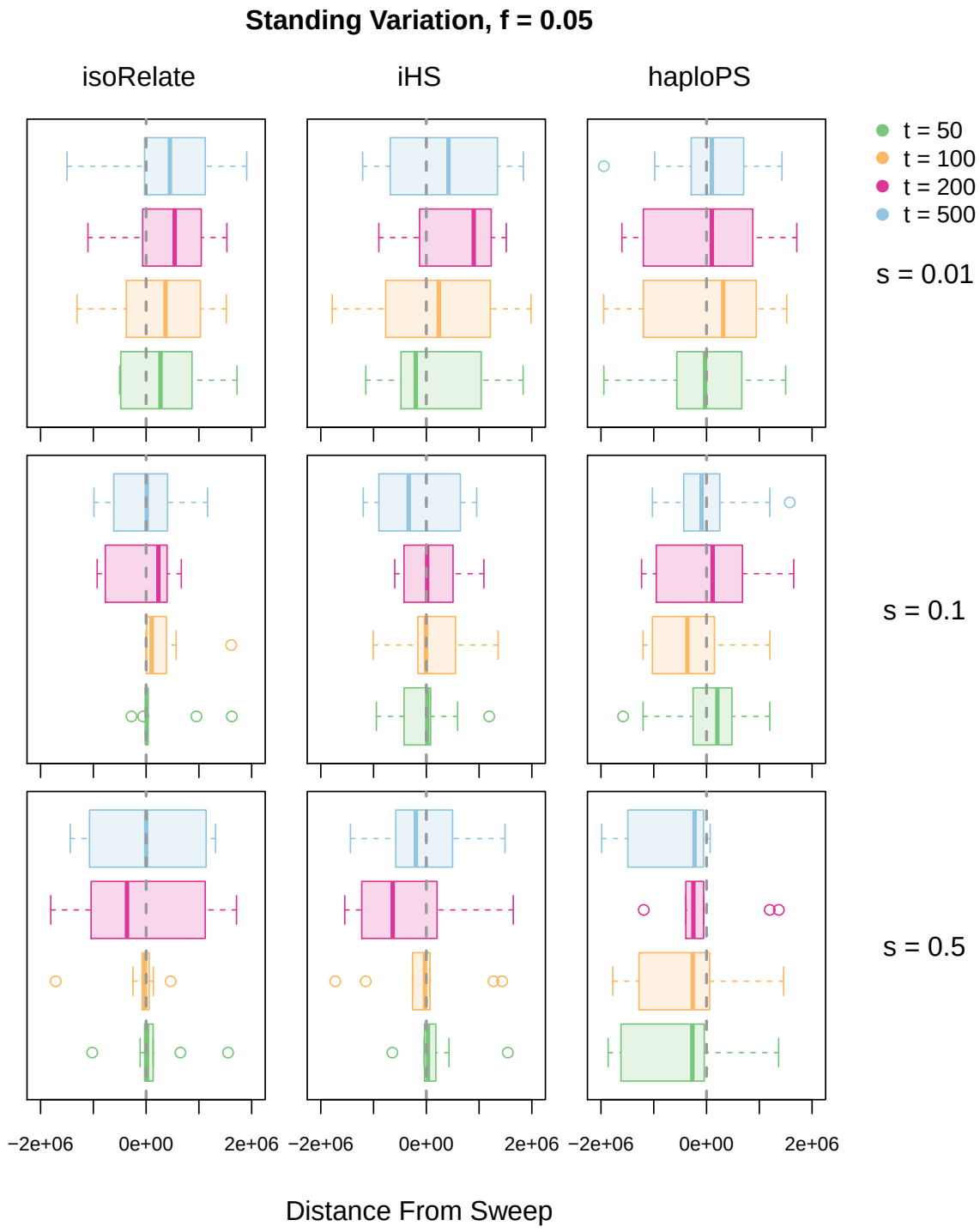


Figure 6.5: Simulation results from standing variation with initial allele frequency $f = 0.05$ for different selection coefficients. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest $-\log_{10}$ p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario.

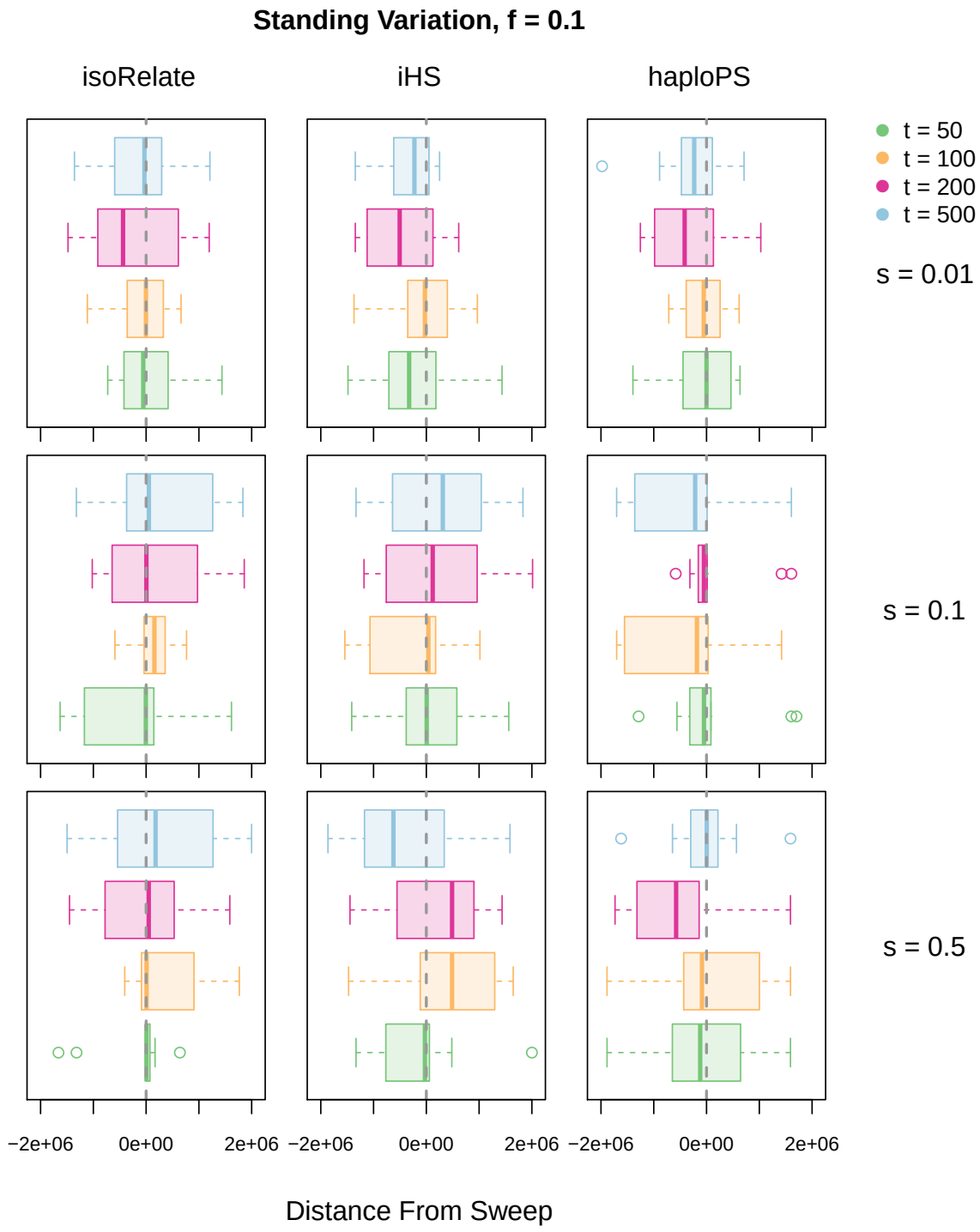


Figure 6.6: Simulation results from standing variation with initial allele frequency $f = 0.1$ for different selection coefficients. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest $-\log_{10}$ p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario.

Multiplicity of infection ≥ 1

isoRelate does not require phased or deconvoluted data, therefore we performed a secondary analysis on 100 isolates with MOI which could exceed 1. Each isolate was assigned MOI according to a zero-truncated Poisson distribution with mean 1. Haplotypes were randomly sampled for each isolate from the 200 haplotypes initially generated for each of the simulation parameter combinations previously examined. Random sampling of haplotypes produces isolates with clonal infections. Both iHS and haploPS were run on only the MOI = 1 isolates with clonal isolates removed while isoRelate was run on all isolates using unphased data.

On average 56% of isolates in each of the 150 datasets have MOI = 1 (Appendix C Table 4). After the removal of clonal isolates, approximately 49 isolates with MOI = 1 remain for analysis with iHS and haploPS in each dataset, while all 100 isolates are used in the analysis of isoRelate. Results for hard sweeps and soft sweeps are comparable to the previous analysis of 200 MOI = 1 isolates, while there is reduced ability to detect selection on standing variation for all three methods (Appendix C Figures 3-7). However, isoRelate detects sweeps that have occurred recently (< 100 generations) more frequently than iHS and haploPS by being able to use the multiclonal isolates.

6.3.3 Population analysis of *P. falciparum*

To demonstrate the ability of isoRelate to investigate a haploid species with known selection signals, we performed IBD mapping of 2,550 P. falciparum isolates from 14 countries across Africa, Southeast Asia and Papua New Guinea as part of the MalariaGEN Pf3K dataset. The samples in this dataset were collected during the years 2001 to 2014 (Appendix C Table 2) and details of the collection process and sequencing protocols have been described elsewhere[97,114]. We define within-country analyses as all pairwise IBD comparisons between isolates from the same country (14 analyses in total) while between-country analyses as all pairwise-country comparisons (91 analyses in total) where pairs of isolates contain one isolate from each country.

After all filtering procedures were complete, 2,377 isolates remained for analysis with 994 isolates (42%) classified as having multiple infections (Appendix C Tables 2 and 5). The mean number of SNPs remaining post filtering for within-country analyses was 31,018 SNPs with the least number of SNPs in the analysis of Papua New Guinea (18,270 SNPs)

and the largest number of SNPs in the analysis of Guinea (44,528 SNPs) (Appendix C Table 2). SNPs for between-country analyses were selected if they appeared in both countries at similar frequencies, which resulted in an average of 12,271 SNPs per analysis with the smallest number of SNPs in the analysis between Mali and Papua New Guinea (1,945 SNPs) and the largest number of SNPs in the analysis between Guinea and Malawi (29,138 SNPs) (Appendix C Table 3). These highly varying numbers of informative SNPs largely reflect geographical isolation and distance but are also influenced by the quality of the WGS data with poorer quality sequencing leading to fewer SNPs. Analyses with so few SNPs, such as Mali and Papua New Guinea, are unlikely to detect selection signatures since smaller IBD segments will fail to be detected, however are still useful for identifying closely related isolates that are expected to share large IBD segments over many SNPs.

6.3.4 Investigating levels of relatedness

We calculated the proportion of pairs IBD at each SNP and investigated the distributions of these statistics across the genome (Figure 6.7, Appendix C Table 6, Appendix C Figure 8). We identified higher levels of relatedness in Southeast Asia than in Africa or in Papua New Guinea, with isolates from Cambodia displaying the highest average sharing across the genome (5%). The Cambodian dataset consists of isolates collected from four study locations; therefore we stratified the relatedness proportions by study location to identify sites with extremely high amounts of relatedness. We detected high relatedness between 87% (2,890/3,321) of pairs from the Pailin Province of Cambodia, with on average 29% of pairs IBD per SNP (Appendix C Tables 7 and 8, Appendix C Figures 9 and 10). Isolates from Pailin contribute 16% of the Cambodian dataset and inflate the overall signal seen in Cambodia. We also detected high amounts of relatedness, including many clonal isolates, in the Thai Province of Sisakhet, which borders Cambodia, reflecting similar transmission dynamics between regions in close proximity.

Relatedness proportions can also be used to identify genomic regions with particularly high amounts of sharing that may be under positive selection as previously shown for IBD studies in human populations[21,22] (Figure 6.7). We observe higher levels of relatedness over several known *P. falciparum* antimalarial drug resistance genes such as *Pfcrt* (chloroquine resistance transporter) and *Pfdhfr* (dihydrofolate reductase) in addition to several regions suspected of being associated with antimalarial drug resistance. In particular, a large proportion of sharing occurs towards the right telomere of chromosome 6, which contains

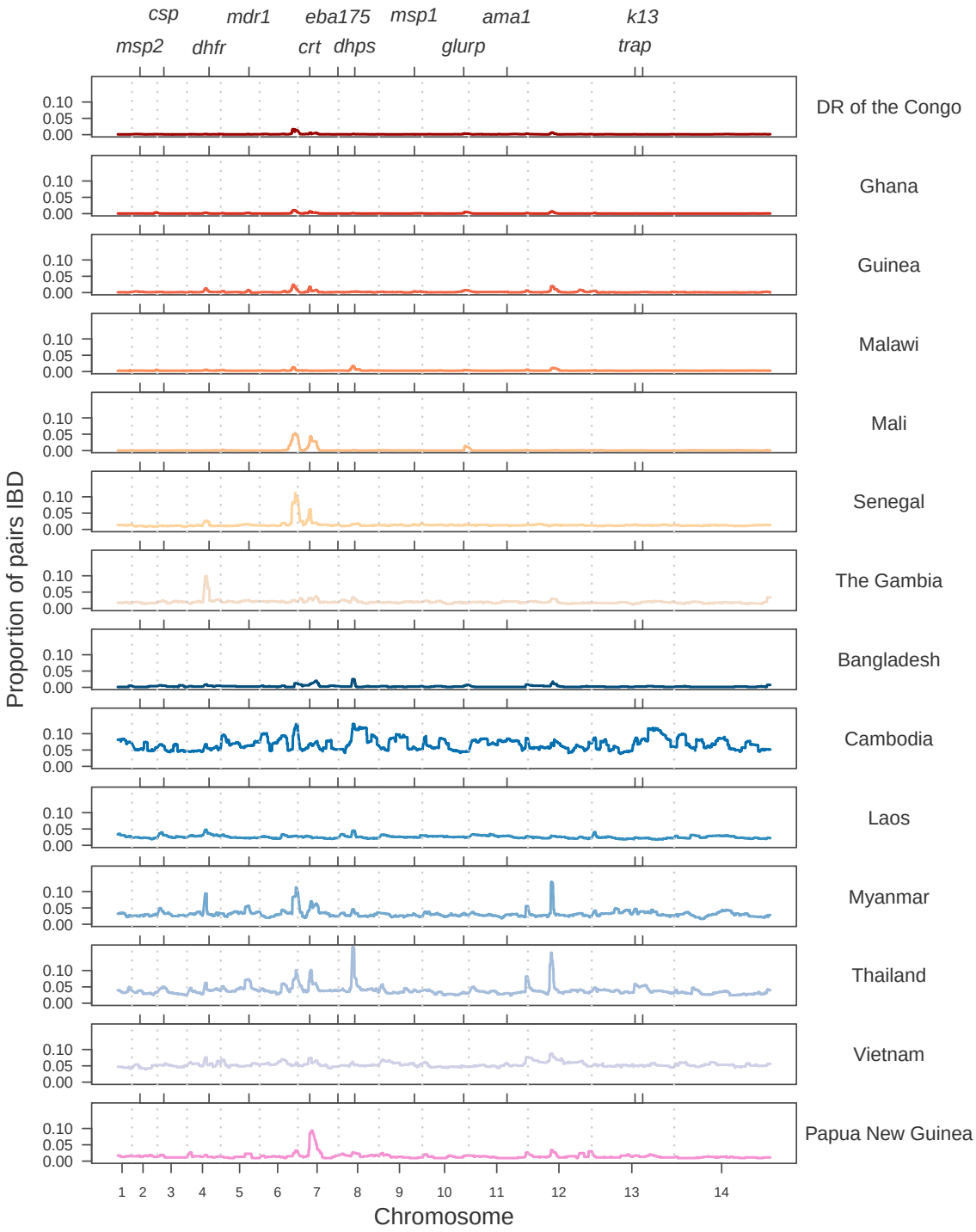


Figure 6.7: The proportion of pairs within each country who are IBD at each SNP, displayed genome wide. Chromosome boundaries are indicated by grey dashed vertical lines and positive control genes are identified gene symbols by tick marks on the top x-axis. Countries that are part of the African continent are shades of red and orange while countries in Southeast Asia are shades of blue and Papua New Guinea is pink.

a number of promising candidate genes suspected of being associated with pyrimethamine resistance[108,125]. Many of these signals also show substantial continent and/or country variation (Figure 6.7).

Relatedness-networks can be created using clustering techniques to identify groups of isolates sharing a common haplotype. We constructed a relatedness-network to investigate clusters of isolates sharing near-identical genomes, reflecting identical infections or 'duplicate' samples (Figure 6.8). Southeast Asia has a number of large clusters containing highly related isolates with the five largest clusters belonging to Cambodia, containing between 12 and 68 isolates, indicative of clonal expansions. The largest cluster contains mostly isolates from the Pursat Province of Cambodia, however the remaining isolates are from the Pailin Province and the Ratanakiri Province of Cambodia, suggesting common haplotypes between western and eastern Cambodia. In contrast, we did not find any isolates within Guinea or Mali to be highly related, nor did we find isolates from different countries to be highly related (Appendix C Table 9, Appendix C Figures 11 - 16).

Some clusters would separate into multiple disjoint clusters if even a single isolate were removed from the group. Isolates which, if removed, would result in disjoint clusters were generally observed to have MOI > 1, where their genome data consists of at least two genetically distinct haplotypes. Such isolates have potentially come from individuals who are traveling between geographical locations and become infected with P. falciparum strains unique to those regions, resulting in IBD that connects multiple, otherwise unconnected sub clusters of isolates.

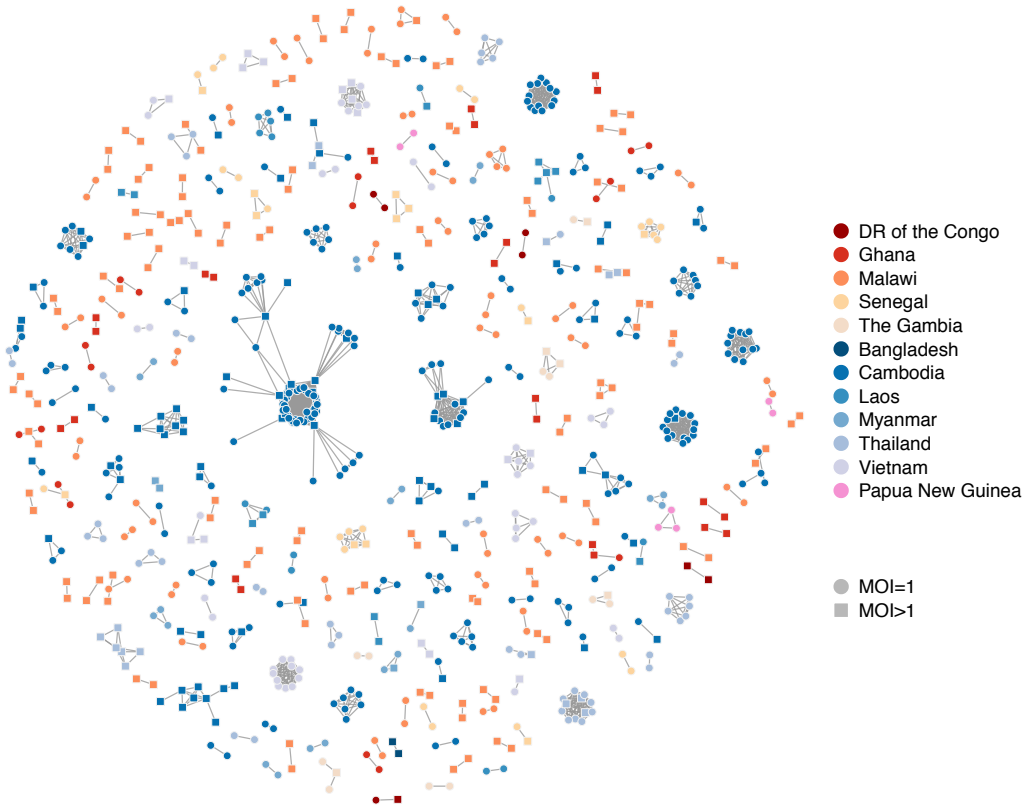


Figure 6.8: Relatedness network showing connections between pairs of isolates identified as having high proportions of IBD sharing. Each node identifies a unique isolates and an edge is drawn between two isolates if they share more than 90% of their genome IBD. Isolates with MOI = 1 are represented by circles while isolates with MOI > 1 are represented by squares. There are 264 clusters in this network comprising 805 isolates (out of 2,377 isolates) in total. Isolates that do not share more than 90% of their genome IBD with any other isolate are omitted from the network.

6.3.5 Analysis of selection signals over the chloroquine resistance locus, *Pfcrt*

To assess the significance of a selection signature we transformed the IBD results for each analysis to account for variations in relatedness between isolates and SNP allele frequencies, then performed normalization allowing us to calculate a new summary metric for each SNP, $-\log 10$ p-values. The genome-wide distributions of the $-\log 10$ p-values for within-country analyses are shown in Figure 6.9 and the top five signals of selection for each country are reported in Appendix C Table 10. We examined in detail the selection signals overlapping the known P. falciparum chloroquine resistance transporter gene, *Pfcrt*, located on chromosome 7 at 403,222-406,317 (Figure 6.10).

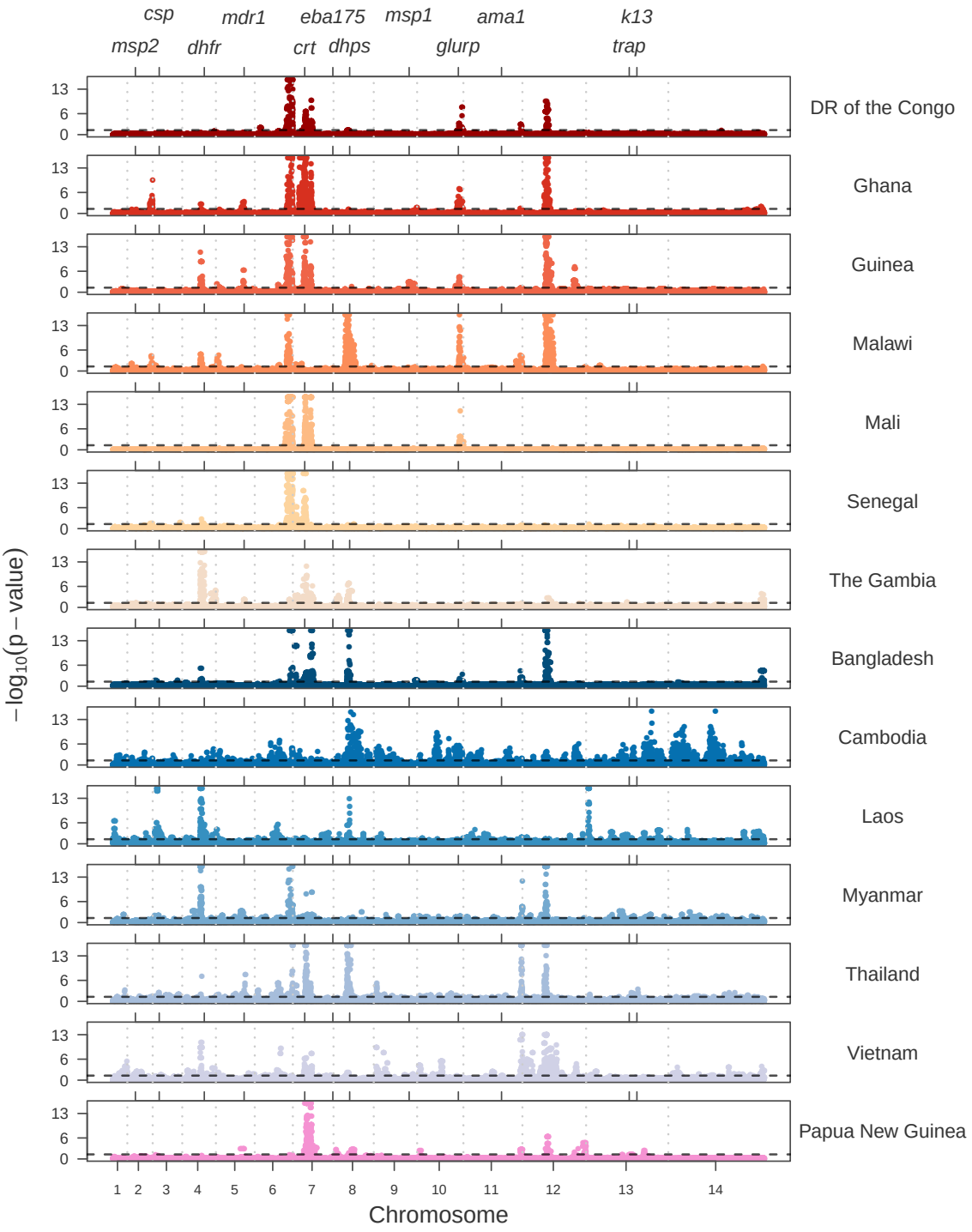


Figure 6.9: $-\log 10$(p-values) of X_{iR} calculated by transforming and normalizing the IBD proportions within each country. Dashed horizontal lines represent a 5% significance threshold. Grey dashed vertical lines indicate chromosome boundaries. Positive control genes are identified by gene symbol and tick marks on the upper x-axis.

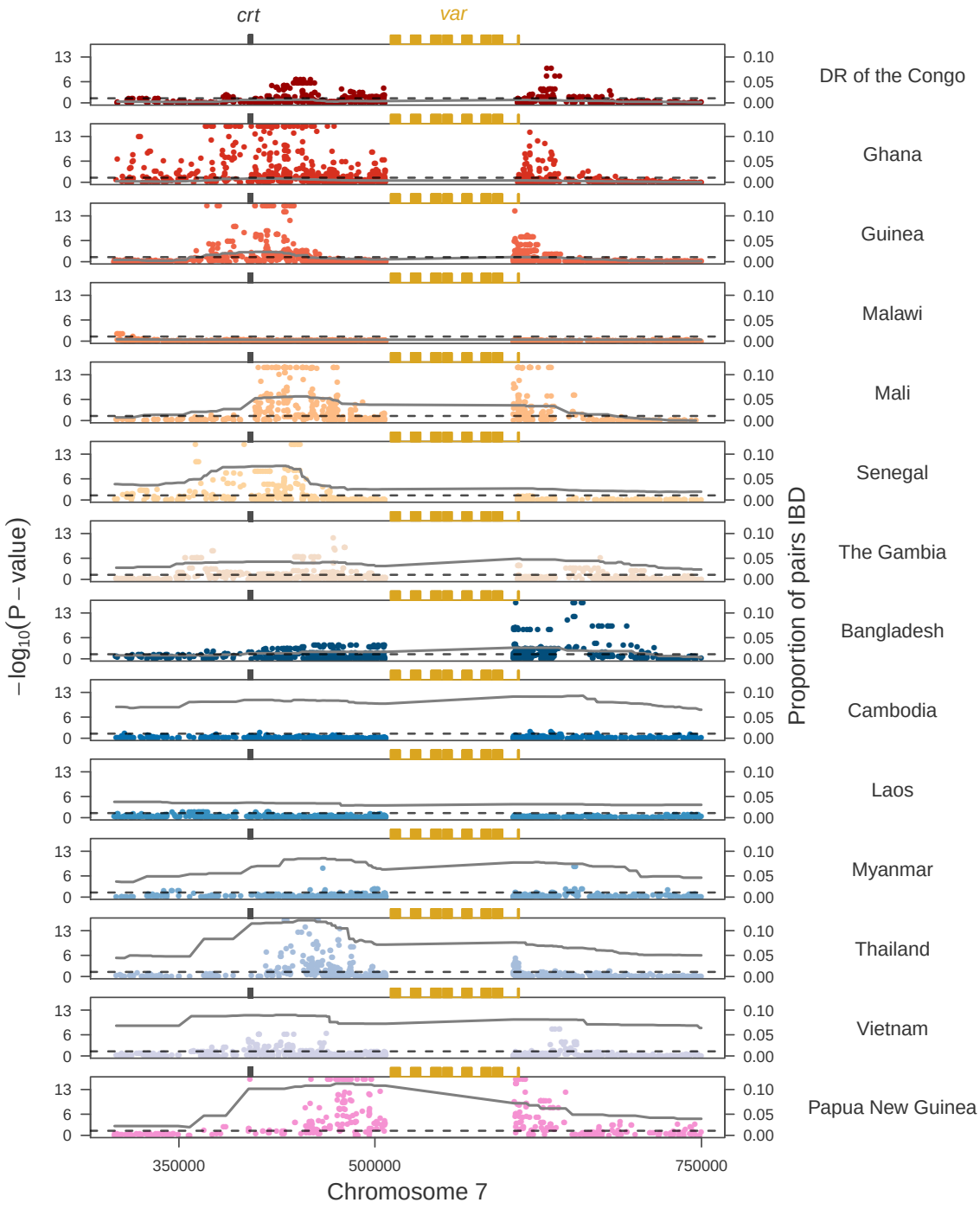


Figure 6.10: Selection signals within each country on chromosome 7 between 300,000-750,000bp, surrounding the *Pfcrt* locus. Coloured points correspond to $-\log 10$(p-values) of the X_{iR} test statistic for all SNPs and the dashed horizontal lines represent the 5% significance threshold. Solid lines are the proportion of pairs IBD over the interval, extrapolated between adjacent SNPs. The *Pfcrt* gene and the *var* gene locations are indicated on the upper x-axis. Note that the *var* gene clusters are blacklisted for the IBD interrogation.

All countries except Malawi and Myanmar have at least one significant SNP within 12kb of *Pfcrt* based on a 5% genome-wide significance threshold. Malawi withdrew the use of chloroquine as an antimalarial drug in 1993, which resulted in the disappearance of the molecular marker of chloroquine resistance (K76T mutation) in Malawian *P. falciparum* populations[126]. Thus we would not expect to see a signature of selection over *Pfcrt* in Malawi. Additionally, none of the between-country analyses involving isolates from Malawi reach significance within 60kb of the *Pfcrt* locus.

Surprisingly, an increase in IBD proportions is observed over *Pfcrt* in Myanmar however the closest significant SNP is located 45kb downstream of *Pfcrt*. In contrast, little to no increase in IBD is observed in the region surrounding *Pfcrt* in Cambodia and Laos, although significant SNPs are identified within close proximity to *Pfcrt*. Both Cambodia and Laos have many isolates sharing large proportions of their genome IBD; potentially adding noise to the summary statistics resulting in inflated significance.

In most countries the highest proportion of IBD on chromosome 7 occurs downstream of *Pfcrt*, with higher levels of IBD extending further downstream of *Pfcrt* than upstream, including over a known set of *var* genes, which were excluded from the IBD analysis due to their complex genetic structure which leads to significant mapping problems. This potentially indicates that *Pfcrt* is regulating a gene downstream or alternatively a second region in close proximity to *Pfcrt* is under selection. A secondary signal immediately downstream of the var genes cluster on chromosome 7 has been previously identified in isolates sampled from The Gambia[127].

We investigated relatedness over *Pfcrt* between isolates from different countries and confirmed the spread of chloroquine resistance throughout Southeast Asia and Africa, while also confirming an independent origin of chloroquine resistance in Papua New Guinea[93,94] (Figure 6.11). However we were unable to determine the exact haplotypes at codons 72-76 of the *Pcfrt* gene, of which CVIET and SVMNT have been associated with chloroquine resistance[93,94], due to low quality data resulting in missing genotype calls for many isolates in addition to unknown haplotype phase for MOI > 1 isolates.

In particular the largest cluster in Figure 3 contains 48% of all isolates, of which 78% have missing genotype calls at codons 73-75 collectively. All isolates in this cluster have the wild type C allele at the C72S variant codon 72. Additionally 95% of these isolates have the chloroquine resistant K76T mutation (codon 76). Thus we speculate the dominant haplotype in the largest cluster to be CVIET, which arose in Southeast Asia

Figure 6.11: Relatedness network showing connections between pairs of isolates inferred IBD over *Pfcrt* (chr7: 403,222-406,317). Each node identifies a unique isolates and an edge is drawn between two isolates if they were inferred IBD anywhere over *Pfcrt*. Isolates with MOI = 1 are represented by circles while isolates with MOI > 1 are represented by squares. There are 178 clusters in this network comprising of 1,563 isolates in total, with the largest cluster containing 1,134 isolates. Isolates that are not IBD over *Pfcrt* are omitted from the network. **A** Isolates are coloured according to country. **B** Isolates are coloured if they carry the K78T mutation associated with chloroquine resistance

and spread to Africa[93]. All isolates from Papua New Guinea have the C72S mutation and K76T mutation (and missing genotype calls at codons 73-75) consistent with the presence of the SVMNT haplotype[94] and these isolates form a separate IBD cluster (bottom of Figure 6.11).

6.3.6 Analysis of selection signals over the artemisinin resistance locus, *Pfk13*

Parasite resistance to the antimalarial drug artemisinin has been associated with mutations in the *P. falciparum* kelch 13 gene, *Pfk13*, located on chromosome 13 at 1,724,817-1,726,997[128,129]. We detected selection signals of marginal significance over *Pfk13* in Cambodia and Thailand (Figure 6.9), which is not surprising given that artermisinin resistance has only recently been identified in Cambodia in 2007 and is currently confined to Southeast Asia[130]. Given the samples from Cambodia and Thailand were collected between 2009 to 2013 (Appendix C Table 2), the resistance mutations are expected to be at low frequencies within these populations, producing very weak signals of selection.

Artemisinin resistance has arisen as a soft selective sweep, involving at least 20 independent *Pfk13* mutations[97]. Relatedness networks over *Pfk13* identify many disjoint clusters of related isolates, with at least 9 clusters containing isolates that carry the most common mutation associated with artemisinin resistance, C580Y[97] (Figure 6.12). We identified isolates from Cambodia, Thailand and Vietnam as carriers of this mutation at frequencies of 40%, 26% and 1% respectively. Additionally, relatedness is detected between isolates from Cambodia and Thailand that have the C580Y mutation as well as isolates from Cambodia and Vietnam with this mutation, suggesting that some resistance-haplotypes have swept between countries[131].

6.3.7 Investigating global inheritance of genomic locations

We investigated the IBD analyses results between countries to determine if any other genomic locations had experienced a global spread like that of chloroquine. We identified a signal on chromosome 6 as having done so, not only between Africa and Southeast Asia, but also Papua New Guinea. In fact, significant IBD sharing is detected in all pairwise-country analyses over the interval chr6: 1,102,005-1,283,312. This interval contains 32 genes of which several have been identified as promising drug resistance candidates[108,125,132]. The cause of this selection pressure remains unknown.

Figure 6.12: Relatedness network showing connections between pairs of isolates inferred IBD over *Pfk13* (chr13: 1,724,817-1,726,997). Each node identifies a unique isolates and an edge is drawn between two isolates if they were inferred IBD anywhere over *Pfk13*. Isolates with MOI = 1 are represented by circles while isolates with MOI > 1 are represented by squares. There are 242 clusters in this network comprising of 1,148 isolates in total, with the largest cluster containing 335 isolates. Isolates that are not IBD over *Pfk13* are omitted from the network. **A** Isolates are coloured according to country. **B** Isolates are coloured if they carry the C580Y mutation associated with artemisinin resistance.

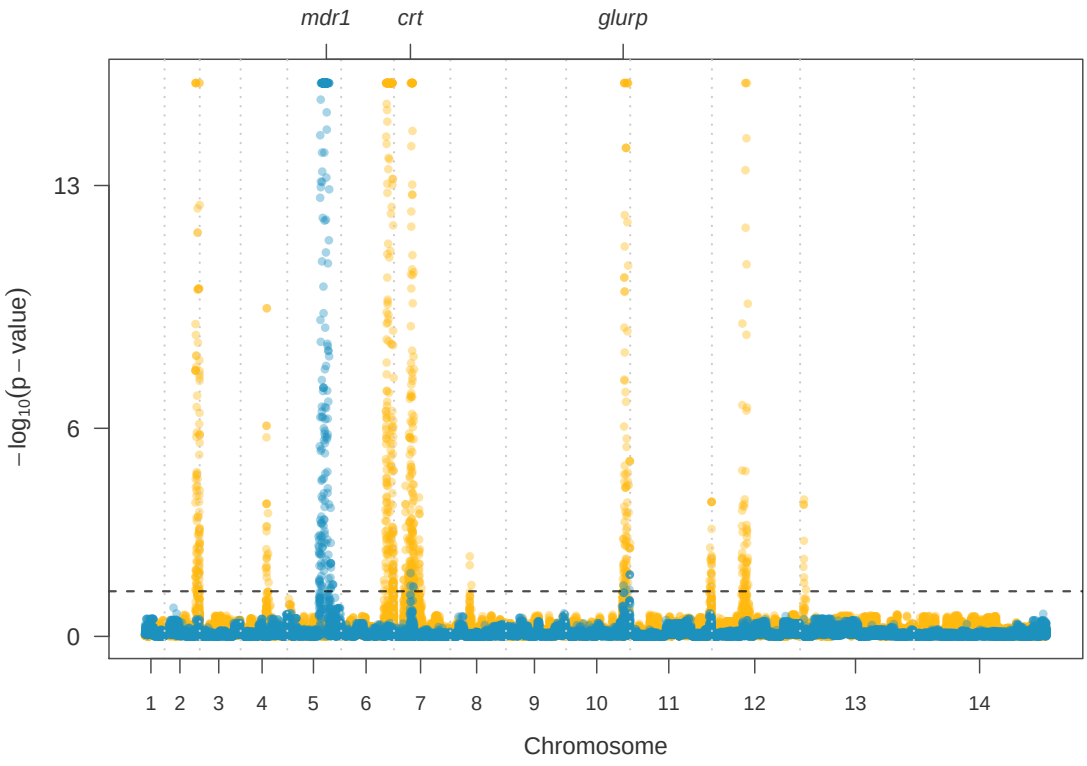


Figure 6.13: Genome-wide selection signatures within Ghana stratified by pairs who are IBD (blue points) or non-IBD (yellow points) over *Pfmdr1* (chr5: 957,890-962,149). The dashed horizontal line represent a 5% significance threshold and the dashed vertical lines identifies the chromosome boundaries. Genomic positions of the genes *Pfmdr1*, *Pfcrt* and *Pfglurp* are indicated by ticks on the upper x-axis.

6.3.8 Detection of multidrug resistance from selection signatures

We explored selection signatures to determine if multidrug resistance could be identified. Specifically, we investigated the *P. falciparum* multidrug resistance gene 1 (*Pfmdr1*), which has been associated with chloroquine resistance and amodiaquine resistance when the *Pfmdr1* N86Y mutation is present along with the *Pfcrt* K76T mutation[133]. Figure 6.13 displays genome-wide selection signals in Ghana, stratified by pairs who are IBD over *Pfmdr1* and pairs who are not IBD over *Pfmdr1*. A significant signal of selection is observed over *Pfcrt* in both stratified groups, suggesting *Pfcrt* is under selection jointly with *Pfmdr1* as well as independently of *Pfmdr1*. Of the isolate pairs who are IBD over *Pfmdr1*, 13% are also IBD over *Pfcrt* while 6% are IBD over *Pfcrt* and carry both the N86Y mutation and the K76T mutation. The median proportion of genome inferred IBD between these pairs is 1%, alleviating concerns that joint inheritance of both variants is due to highly related pairs. An additional selection signal is identified over *Pfglurp* (glutamine-rich protein, a candidate vaccine antigen) also suggesting joint selection of both *Pfmdr1*

and *Pfglurp*.

6.3.9 Analysis of selection signal methodologies on global *P. falciparum* dataset

We compared the selection signatures generated by isoRelate within countries to those detected by iHS[101]. The EHH algorithm requires knowledge of haplotype phase, which is currently not possible for isolates with MOI > 1 as the number of strains in an infection and the proportions they contribute to the mixed infection must be known, in addition to having quality data sequenced at high coverage. In contrast, haplotype phase is trivial for isolates with MOI $= 1$, therefore we performed comparisons of isoRelate and iHS using only isolates with MOI $= 1$, on the same SNPs (Appendix C Table 11, Appendix C Figure 17). The largest $-\log 10$ p-value for a single SNP within each of 12 interesting genes is reported in Appendix C Table 12.

Although there is some overlap in the selection signatures produced by iHS and isoRelate, there is a surprising dissimilarity between the results. iHS detects selection at *Pfglurp*, *Pfama1* and *Pftrap* more frequently than isoRelate, however has difficulty detecting selection in Southeast Asia and Papua New Guinea. In contrast isoRelate commonly detects selection over *Pfdhfr*, *Pfmdr1*, *Pfcrt* and *Pfdhps*. Additionally prominent signals are also detected on chromosome 6 and chromosome 12 by isoRelate, in regions that, as yet, have no reported candidate genes.

The genes *Pfglurp*, *Pfama1* and *Pftrap* detected by iHS encode surface proteins that undergo balancing selection[134,135] and hence have been investigated as vaccine targets. We anticipate selection on extremely recent mutations in these genes that are at low frequency within the population, in which case iHS is more likely to detect this selection than isoRelate. This is simply because iHS profiles are calculated relative to the number of isolates in a population while isoRelate profiles are calculated relative to the number of pairwise combinations in the population, which heavily dilutes excess IBD sharing of low frequency haplotypes.

Additionally, iHS assumes that all samples are independent, meaning there is no relatedness between isolates. This assumption is violated in all countries, particularly in Southeast Asia where there are many highly related isolates, preventing iHS from decaying to a threshold at some SNPs, resulting in missing iHS values. On average 84% of SNPs in African countries have missing iHS values, while 94% of SNPs in Southeast Asian

countries have missing values, contributing to the lack of signals detected in Asia. To avoid such loss of information, related isolates could be removed from iHS analyses at the risk of reduced power due to smaller sample sizes. However in some instances the sample size would reduce significantly, as is the case with Cambodia, which would experience an 80% reduction in sample size if isolates sharing more than 10% of their genome IBD were removed. Considering the number of SNPs with missing values, iHS does surprisingly well in African countries.

6.4 Discussion

Relatedness mapping of microorganisms is useful for investigating the genetic mechanisms involved in diseases. We demonstrate this on a global whole-genome sequenced *P. falciparum* dataset using a new IBD method, isoRelate, which provides novel insights into the geographical spread of antimalarial drug resistance, including multidrug resistance, as well as population structure.

IBD inference of *P. falciparum* genomes allows us to compare different levels of relatedness between geographical regions. Here we identified the Pailin Province of Cambodia as having many highly related isolates, either as a result of intensified malaria control efforts following the emergence of artemisinin resistance in 2007[130] or as an artifact of the sampling collection procedures, in which case greater efforts may need to be made to attain independence for population genetic studies. As such, we propose genome wide IBD summaries as a means of monitoring malaria control, whereby intensified control regimes reduce malaria transmission and genetic diversity[136,137], resulting in more relatedness between strains and higher proportions of IBD.

Our algorithm allows us to infer IBD status at any genomic location, which lead us to develop a new summary measure of IBD sharing in populations at genomic locations, resulting in a novel measure for detecting selection. We developed a statistical framework to test the significance of selection signatures, which, unlike iHS, accounts for the level of relatedness between isolates. Using the IBD approach we were able to identify both known resistance loci, underpinned by known resistance genes, including *Pfcrt* and *Pfk13*, and several novel signals of selection, one of which has been previously reported on chromosome 6[108,125,132]. Quantifying relatedness is important in analyses wishing to investigate selection, as highly related isolates add noise to the results, making it harder

to identify selection signatures. Ideally related isolates would be excluded from analyses, however as disease control reduces transmission, highly related isolates will become prominent (Daniels et al. 2015) and removing these isolates could greatly reduce the power of the analysis.

We generated relatedness networks to provide insights into the number of haplotypes within a genomic interval as well as their origin, which has immediate applications for monitoring the geographic spread of antimicrobial drug resistant haplotypes. We visualized the spread of chloroquine resistance across Southeast Asia and Africa using such networks, confirming an independent origin of resistance in Papua New Guinea[93,94]. We also examined relatedness over *Pfk13* and were able to visualize a number of founder haplotypes carrying the C580Y mutation, associated with artemisinin resistance, also confirming that resistance to artemisinin has arisen as a soft selective sweep[97].

IBD analyses require several criteria to be met. This includes the availability of a good quality reference genome and the fact that the organism must recombine as one of its main sources of creating genetic variation. As such these methods do not appear to be applicable to *Mycobacterium tuberculosis* for example, but will work with any other organism that shares these criteria with *P. falciparum*. Amongst these are *P. vivax*[138] and some species of *Staphylococcus*[139]. Thus isoRelate will have broader application than just *P. falciparum*. Furthermore, isoRelate can be applied to any dense genomic data that produces SNP genotypes, which includes WGS, RNA sequencing and SNP arrays.

isoRelate is the first algorithm to implement an IBD-based selection detection approach applicable for field isolates with possible multi-clonality. We have shown that our approach can dissect complex signals of selection, including selection on standing variation. This method will be invaluable for the identification and genomic surveillance of drug resistance loci in many microorganisms.

Chapter 7

Discussion and conclusion

7.1 Summary

IBD analysis is a valuable tool with a plethora of applications, including identification of candidate disease genes, detecting unknown relatedness and identifying selection signatures. Analysis is commonly performed on autosomal chromosomes as methodologies are typically developed for diploid genomes. As such, the X chromosome is generally excluded from analysis as it requires special treatment to account for differences in chromosomal numbers between males and females. Exclusion of the X chromosome is not only common in IBD analysis but also linkage analysis and GWAS, resulting in much fewer discoveries on the X chromosome relative to other chromosomes[140]. This is unfortunate as evidence suggests that the X chromosome plays an important role in human disease[140,141,142]. The lack of IBD methodologies for the X chromosome not only affects human analysis, but analysis of organisms with haploid genomes, such as the malaria-causing parasite *Plasmodium* and bacterium *Mycobacterium tuberculosis*, limiting our understanding of other diseases.

This thesis describes a novel methodology for IBD analysis of both diploid and haploid chromosomes in organisms that undergo recombination, with both human and non-human applications. Chapter 2 introduces the methodology XIBD, which is an extension of the HMM developed by Purcell et al.[41] and Albrechtsen et al.[18]. The methodology described in this chapter was published in Bioinformatics in 2016[63].

Chapter 3 evaluates the performance of XIBD on a simulated dataset, varying LD with different ploidies, demonstrating improved IBD performance as haploid chromosomes are included in the analysis and identifying an LD filtering threshold of $R^2 \geq 0.8$. We also

perform comparisons between XIBD, GERMLINE and fastIBD and determine that XIBD performs similarly, or otherwise outperforms these tools.

In Chapter 4 we present the results of an analysis using XIBD on a cohort of 11 families with a rare form of epilepsy. We discover unknown-relatedness between four families; are able to refine the disease critical region; and provide evidence of a founder effect, with at least four distinct founders responsible for the disorder. This information can help in the identification of causal variants as it constrains the expectations regarding the number of variants and who should carry them. The findings from this chapter were published in Human Genetics in 2016[143].

Following the description of our methodology in Chapter 3 and application to an epilepsy cohort in Chapter 4, we introduce the burdensome disease malaria and the parasite responsible for malaria, *Plasmodium*. In Chapter 5 we discuss the need for an IBD methodology that is applicable to haploid organisms, like *Plasmodium*, that can be used to identify loci under positive selection. This is motivated by the emergence of antimalarial drug resistance and the importance of identifying the genetic mechanisms underpinning such resistance.

Chapter 6 then details how the IBD model described in Chapter 2 can be applied to isolates with single or multiple infections of haploid microorganisms. Here, we perform an IBD analysis on a global *P. falciparum* dataset, the deadliest species of *Plasmodium* that infects humans, and demonstrate its usefulness in a number of applications, most importantly in determining loci under positive selection. We identify previously known loci under selection due to antimalarial drug pressure, including *Pfcrt* associated with chloroquine resistance, in addition to several loci suspected of being associated with antimalarial drug resistance. Furthermore, we confirm the spread of chloroquine resistance throughout Southeast Asia and Africa; confirm independent origins of resistance to artemisinin; and infer fine-scale population structure within and between countries. The work presented in Chapter 6 is being prepared for submission to Molecular Biology and Evolution. A preprint is available on BioRxiv[112].

7.2 Importance and implications of the methodology and results

The methodology presented in Chapter 2 allows for novel insights into relatedness on the X chromosome, which is often neglected in analyses. This method will be extremely useful for disorders that have an X-linked component, which, according to the Online Mendelian Inheritance in Man, account for 6.4% of phenotypes with a known molecular basis and includes disorders such as epilepsy, autism and intellectual disability[141]. The development of an IBD methodology for the X chromosome should pave the way for a greater number of discoveries of X-linked genes causal for these, and other, disorders.

The work presented in Chapters 5 and 6 highlights the importance of applying IBD analysis to organisms other than humans. Following identification of two novel loci under positive selection in *P. falciparum*, on chromosomes 6 and 12 respectively, molecular analyses are now underway to identify any associations with antimalarial drug resistance. In particular, the locus on chromosome 12, which was identified in 11 countries throughout Africa, Southeast Asia and Papua New Guinea (PNG), is under investigation. Approximately 20 field isolates from PNG, some of which were included in our analysis, have been cultured and are undergoing screening to determine which antimalarial drug(s), if any, the isolates have reduced sensitivity towards. If drug screening successfully identifies reduced sensitivity, a list of candidate variants extracted from the PNG WGS data within the locus on chromosome 12 will be produced. CRISPR knockout will then be performed to confirm which variants are associated with antimalarial drug resistance.

The IBD methodology described in Chapter 6 can also be applied to species other than *P. falciparum*. Future analyses will include longitudinal studies of *P. vivax*, which can remain dormant within the human liver for months or years after the initial infection before resurfacing, to investigate whether an infection is recurrent or not. An infection can be classified as recurrent if a large amount of relatedness in inferred with previous infections. Bright et al.[144] performed a similar analysis using IBS to confirm that three relapse infections in a patient from Africa were all caused by meiotic siblings that were the result of a single meiosis event within a mosquito.

An additional analysis that is beyond the scope of this thesis, is to explore IBD within the highly-polymorphic *var* gene family. The *var* genes encode hypervariable surface proteins that are crucial for *P. falciparum* to evade the human immune response to infec-

tion[145]. A number of conserved *var* genes have been associated with chloroquine resistance as a result of their physical linkage to *Pfcrt*[146]. Additional insights into the mechanisms of drug resistance could be gained if more conserved *var* genes are identified that are associated with drug resistance. However, identifying conservation, or IBD, over the *var* genes requires *de novo* assembly of sequenced data, which is difficult because of the high AT content of *P.falciparum*[147], in addition to the high polymorphism resulting from complex recombination processes whereby *var* genes undergo ectopic recombination[90] in addition to recombination during both meiosis and mitosis within the mosquito and human hosts, respectively[145]. It is unclear whether IBD approaches will be successful given the complexity of the *var* genes and other avenues may need to be explored to determine *var* gene conservation.

An interesting extension of IBD analysis in *Plasmodium* is to determine IBD between strains *within* an infection, i.e., detect IBD between strains within an isolate with MOI > 1. We are currently developing a model to do exactly this using B-allele frequency data (i.e., the proportion of reads that map to the alternative allele at a locus), similar to models used to infer copy number variation in human studies. This model is likely to be complex and presents some challenges due to the AT bias of the *P. falciparum* genome, the unknown number of strains in an infection and the relative contribution of each strain to the infection. Furthermore, such a model may be redundant with the recent development of DEpolid[86], a deconvolution method that estimates the number of strains in an infection, the relative contribution of each strain to the infection and the haplotype of each strain. With haplotype data available, IBD analysis as described in Chapter 2 could be applied directly to this data, eliminating the need for a second model. DEploid has the potential to revolutionize genomic analyses of *Plasmodium*, and organisms with similar properties to *Plasmodium*, advancing the wealth and quality of knowledge of a number of diseases.

In conclusion, this PhD thesis describes the successful development and application of an IBD methodology for diploid and haploid genomes, for both human and non-human applications. It has provided new insights into the rare epilepsy disorder, FAME, as well as the deadliest malaria-causing parasite, *P. falciparum*. We anticipate our methodology to become more widely used as advances in technologies allow for rapid sequencing of many more microorganisms.

Bibliography

[1] Anthony J K Griffiths, Susan R Wessler, Richard C Lewontin, and Sean B Carroll. Introduction to genetic analysis, 9th edition. England, 2008.

[2] Brian Charlesworth. Sex Determination in the Honeybee. *Cell*, 114(4):397–398, August 2003.

[3] Yannick Wurm, John Wang, Oksana Riba-Grognuz, Miguel Corona, Sanne Nygaard, Brendan G Hunt, et al. The genome of the fire ant Solenopsis invicta. *PNAS*, 2011.

[4] S Godreuil, L Tazi, and A L Bañuls. Pulmonary tuberculosis and Mycobacterium tuberculosis: modern molecular epidemiology and perspectives. *Encyclopedia of infectious diseases: modern methodologies*, 2007.

[5] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511, October 2002.

[6] A Annunziato. DNA Packaging: Nucleosomes and Chromatin — World Library of Science. *Nature Education*, 2008.

[7] J Ott. Analysis of human genetic linkage. Baltimore, 1999.

[8] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[9] Eric S Lander, Michele Cargill, David Altshuler, James Ireland, Pamela Sklar, Kristin Ardlie, Nila Patil, Charles R Lane, Esther P Lim, Nilesh Kalyanaraman, James Nemesh, Liuda Ziaugra, Lisa Friedland, Alex Rolfe, Janet Warrington, Robert

Lipshutz, and George Q Daley. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22(3):231–238, July 1999.

[10] D J Witherspoon, S Wooding, A R Rogers, E E Marchani, W S Watkins, M A Batzer, and L B Jorde. Genetic Similarities Within and Between Human Populations. *Genetics*, 176(1):351–359, February 2007.

[11] J C Venter. The Sequence of the Human Genome. *Science*, 291(5507):1304–1351, February 2001.

[12] A Helena Mangs and Brian J Morris. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. *Current genomics*, 8(2):129–136, April 2007.

[13] Eric L Stevens, Greg Heckenberg, Elisha D O Roberson, Joseph D Baugher, Thomas J Downey, and Jonathan Pevsner. Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS genetics*, 7(9):e1002287, September 2011.

[14] Alun Thomas, Mark H Skolnick, and Cathryn M Lewis. Genomic mismatch scanning in pedigrees. *Mathematical Medicine and Biology*, 11(1):1–16, 1994.

[15] Sharon R Browning and Brian L Browning. Identity by descent between distant relatives: detection and applications. *Annual review of genetics*, 46:617–633, 2012.

[16] Elizabeth A Thompson. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, June 2013.

[17] M D Brown, C G Glazner, C Zheng, and E A Thompson. Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*, 190(4): 1447–1460, April 2012.

[18] Anders Albrechtsen, Thorfinn Sand Korneliussen, Ida Moltke, Thomas van Overseem Hansen, Finn Cilius Nielsen, and Rasmus Nielsen. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology*, 33(3):266–274, April 2009.

[19] Marie Shaw, Tzu Ying Yap, Lyndal Henden, Melanie Bahlo, Alison Gardner, Vera M Kalscheuer, Eric Haan, Louise Christie, Anna Hackett, and Jozef Gecz. Identical by descent L1CAM mutation in two apparently unrelated families with intellectual

disability without L1 syndrome. *European journal of medical genetics*, 58(6-7):364–368, June 2015.

[20] Trevor J Pemberton, Chaolong Wang, Jun Z Li, and Noah A Rosenberg. Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *American journal of human genetics*, 87(4):457–464, October 2010.

[21] Anders Albrechtsen, Ida Moltke, and Rasmus Nielsen. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308, September 2010.

[22] Lide Han and Mark Abney. Using identity by descent estimation with dense genotype data to detect positive selection. *European journal of human genetics : EJHG*, 21 (2):205–211, February 2013.

[23] M Gunay-Aygun, Y Zivony-Elboum, F Gumruk, D Geiger, M Cetin, M Khayat, R Kleta, et al. Gray platelet syndrome: natural history of a large patient cohort and locus assignment to chromosome 3p. *Blood*, 116(23):4990–5001, December 2010.

[24] Margaret A Pericak-Vance. *Analysis of Genetic Linkage Data for Mendelian Traits*. John Wiley & Sons, Inc., Hoboken, NJ, USA, May 2001.

[25] David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33 Suppl:228–237, March 2003.

[26] Cornelis A Albers, Jim Stankovich, Russell Thomson, Melanie Bahlo, and Hilbert J Kappen. Multipoint approximations of identity-by-descent probabilities for accurate linkage analysis of distantly related individuals. *American journal of human genetics*, 82(3):607–622, March 2008.

[27] Sharon R Browning and Brian L Browning. High-resolution detection of identity by descent in unrelated individuals. *American journal of human genetics*, 86(4): 526–539, April 2010.

[28] Roderick H J Houwen, Siamak Baharloo, Kathleen Blankenship, Peter Raeymaekers, Jenneke Juyn, Lodewijk A Sandkuijl, and Nelson B Freimer. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genetics*, 8(4):380–386, December 1994.

[29] P M Visscher, M A Brown, and M I McCarthy. Five Years of GWAS Discovery. *The American Journal of . . .*, 2012.

[30] Sharon R Browning and Elizabeth A Thompson. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–1531, April 2012.

[31] Rui Lin, Jac Charlesworth, Jim Stankovich, Victoria M Perreau, Matthew A Brown, ANZgene Consortium, and Bruce V Taylor. Identity-by-descent mapping to detect rare variants conferring susceptibility to multiple sclerosis. *PloS one*, 8(3):e56379, 2013.

[32] William S Bush and Jason H Moore. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[33] Augustine Kong, Gisli Masson, Michael L Frigge, Arnaldur Gylfason, Pasha Zusmanovich, Gudmar Thorleifsson, Pall I Olason, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, 40(9):1068–1075, September 2008.

[34] Joseph J Vitti, Sharon R Grossman, and Pardis C Sabeti. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, 47(1):97–120, November 2013.

[35] Laura B Scheinfeldt and Sarah A Tishkoff. Recent human adaptation: genomic approaches, interpretation and insights. *Nature reviews. Genetics*, 14(10):692–702, October 2013.

[36] Jonathan K Pritchard, Joseph K Pickrell, and Graham Coop. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology : CB*, 20(4):R208–15, February 2010.

[37] Thomas LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, 37(13): 4181–4193, July 2009.

[38] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, June 2011.

[39] Ian W Saunders, Jesper Brohede, and Garry N Hannan. Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. *Genomics*, 90(3):291–296, September 2007.

[40] Gonçalo R Abecasis, Stacey S Cherny, William O Cookson, and Lon R Cardon. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1):97–101, December 2001.

[41] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81 (3):559–575, September 2007.

[42] S Bercovici, C Meek, Y Wexler, and D Geiger. Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. *Bioinformatics*, 26(12):i175–i182, June 2010.

[43] Peter M Krawitz, Michal R Schweiger, Christian Rödelsperger, Carlo Marcelis, Uwe Kölsch, Christian Meisel, Friederike Stephani, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nature Genetics*, 42(10):827–829, October 2010.

[44] Lide Han and Mark Abney. Identity by descent estimation with dense genome-wide genotype data. *Genetic Epidemiology*, pages n/a–n/a, July 2011.

[45] Dan He. IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. *Bioinformatics*, 29(13):i162–70, July 2013.

[46] Chris Glazner and Elizabeth Thompson. Pedigree-Free Descent-Based Gene Mapping from Population Samples. *Human heredity*, 80(1):21–35, 2015.

[47] Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326, February 2009.

[48] Ian M Carr, Eamonn Sheridan, Bruce E Hayward, Alexander F Markham, and David T Bonthron. IBDfinder and SNPsetter: tools for pedigree-independent identi-

fication of autozygous regions in individuals with recessive inherited disease. *Human Mutation*, 30(6):960–967, June 2009.

[49] Sharon R Browning and Brian L Browning. A fast, powerful method for detecting identity by descent. *American journal of human genetics*, 82(2):173–182, February 2011.

[50] A Jacquard. The Genetic Structure of Populations. New York, 1974.

[51] I Moltke, A Albrechtsen, T v O Hansen, F C Nielsen, and R Nielsen. A method for detecting IBD regions simultaneously in multiple individuals–with applications to disease genetics. *Genome Research*, 21(7):1168–1180, July 2011.

[52] Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, June 2013.

[53] Brian L Browning and Sharon R Browning. Detecting identity by descent and estimating genotype error rates in sequence data. *American journal of human genetics*, 93(5):840–851, November 2013.

[54] Sepp Hochreiter. HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data. *Nucleic Acids Research*, 41(22):e202, December 2013.

[55] Jesse M Rodriguez, Sivan Bercovici, Lin Huang, Roy Frostig, and Serafim Batzoglou. Parente2: A fast and accurate method for detecting identity by descent. *Genome Research*, October 2014.

[56] Wenqing Fu, Sharon R Browning, Brian L Browning, and Joshua M Akey. AR TICLERobust Inference of Identity by Descent from Exome-Sequencing Data. *American journal of human genetics*, pages 1–11, October 2016.

[57] Hitoshi Miyazawa, Masaaki Kato, Takuya Awata, Masakazu Kohda, Hiroyasu Iwasa, Nobuyuki Koyama, Tomoaki Tanaka, Huqun, Shunei Kyo, Yasushi Okazaki, and Koichi Hagiwara. Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *American journal of human genetics*, 80 (6):1090–1102, June 2007.

[58] A Thomas, N J Camp, J M Farnham, K Allen Brady, and L A Cannon Albright. Shared Genomic Segment Analysis. Mapping Disease Predisposition Genes in Extended Pedigrees Using SNP Genotype Assays. *Annals of Human Genetics*, 72(2): 279–287, 2008.

[59] Alun Thomas. Assessment of SNP streak statistics using gene drop simulation with linkage disequilibrium. *Genetic epidemiology*, 34(2):119–124, February 2010.

[60] Brian L Browning and Sharon R Browning. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology*, 31(5):365–375, July 2007.

[61] S R Browning and B L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 2007.

[62] S R Browning. Estimation of Pairwise Identity by Descent From Dense Genetic Marker Data in a Population Sample of Haplotypes. *Genetics*, 178(4):2123–2132, April 2008.

[63] L Henden, D Wakeham, and M Bahlo. XIBD: software for inferring pairwise identity by descent on the X chromosome. *Bioinformatics*, 32(15):2389–2391, 2016.

[64] L R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[65] Yingjian Zhang. *Prediction on financial time series with midden Markov models*. PhD thesis, Simon Fraser University, 2001.

[66] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F A Grant, Hakon Hakonarson, and Maja Bucan. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome ...*, 2007.

[67] F A Sonnenberg and J R Beck. Markov models in medical decision making: a practical guide. *Medical decision making : an international journal of the Society for Medical Decision Making*, 13(4):322–338, October 1993.

[68] M S McPeek and L Sun. Statistical tests for detection of misspecified relationships by use of genome-screen data. *American journal of human genetics*, 66(3):1076–1094, March 2000.

[69] M P Epstein, W L Duren, and M Boehnke. Improved inference of relationship for pairs of individuals. *American journal of human genetics*, 67(5):1219–1231, November 2000.

[70] Mark Abney. A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics*, 25(12):1561–1563, June 2009.

[71] JBS Haldane. The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, 1919.

[72] D Clayton and H T Leung. An R package for analysis of whole-genome association studies. *Human heredity*, 2007.

[73] The International HapPap Consortium. The International HapMap Project. *Nature*, 426(6968):789–796, December 2003.

[74] M Bahlo and C J Bromhead. Generating linkage mapping files from Affymetrix SNP chip data. *Bioinformatics (Oxford, England)*, 25(15):1961–1962, August 2009.

[75] Alexander Gusev, Eimear E Kenny, Jennifer K Lowe, Jaqueline Salit, Richa Saxena, Sekar Kathiresan, David M Altshuler, Jeffrey M Friedman, Jan L Breslow, and Itsik Pe'er. DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *American journal of human genetics*, 88(6):706–717, June 2011.

[76] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL `http://ggplot2.org`.

[77] Katherine R Smith, Catherine J Bromhead, Michael S Hildebrand, A Eliot Shearer, Paul J Lockhart, Hossein Najmabadi, Richard J Leventer, George McGillivray, David J Amor, Richard J Smith, and Melanie Bahlo. Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biology*, 12(9):R85, 2011.

[78] Saskia Freytag, Johann Gagnon-Bartsch, Terence P Speed, and Melanie Bahlo. Systematic noise degrades gene co-expression signals but can be corrected. *BMC bioinformatics*, 16:309, September 2015.

[79] World Health Organization. Fact sheet: world malaria day 2016, 2016. URL `http://www.who.int/malaria/media/world-malaria-day-2016/en/`.

[80] A F Cowman, J Healer, D Marapana, and K Marsh. Malaria: Biology and Disease. *Cell*, 2016.

[81] World Health Organization. Fact sheet: world malaria report 2015, 2015. URL `http://www.who.int/malaria/media/world-malaria-report-2015/en/`.

[82] World Health Organization. Antimicrobial resistance, 2016. URL `http://www.who.int/mediacentre/factsheets/fs194/en/`.

[83] E Y Klein. Antimalarial drug resistance: a review of the biology and strategies to delay emergence and spread. *International Journal of Antimicrobial Agents*, 41(4): 311–317, April 2013.

[84] Sarah Auburn, Susana Campino, Taane G Clark, Abdoulaye A Djimde, Issaka Zongo, Robert Pinches, Magnus Manske, et al. An effective method to purify Plasmodium falciparum DNA directly from clinical blood samples for whole genome high-throughput sequencing. *PloS one*, 6(7):e22213, 2011.

[85] P A Zimmerman, R K Mehlotra, and L J Kasehagen. Why do we need to know more about mixed Plasmodium species infections in humans? *Trends in ...*, 2004.

[86] Sha Joe Zhu, Jacob Almagro-Garcia, and Gil McVean. Deconvoluting multiple infections in Plasmodium falciparumfrom high throughput sequencing data. *bioRxiv*, 2017.

[87] Kevin Galinsky, Clarissa Valim, Arielle Salmier, Benoit de Thoisy, Lise Musset, Eric Legrand, et al. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malaria journal*, 14:4, January 2015.

[88] John D O'Brien, Zamin Iqbal, Jason Wendler, and Lucas Amenga-Etego. Inferring Strain Mixture within Clinical Plasmodium falciparum Isolates from Genomic Sequence Data. *PLoS computational biology*, 12(6):e1004824, June 2016.

[89] Alistair Miles, Zamin Iqbal, Paul Vauterin, Richard Pearson, Susana Campino, Michel Theron, Kelda Gould, et al. Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome research*, 26(9): 1288–1299, September 2016.

[90] Sue A Kyes, Susan M Kraemer, and Joseph D Smith. Antigenic variation in Plasmodium falciparum: gene organization and regulation of the var multigene family. *Eukaryotic cell*, 6(9):1511–1520, September 2007.

[91] Srinivas. Evolution of malaria parasites, 2015. URL `http://www.malariasite.com/history-parasites/`.

[92] Thomas E Wellems and Christopher V Plowe. Chloroquine-Resistant Malaria. *The Journal of Infectious Diseases*, 184(6):770–776, September 2001.

[93] John C Wootton, Xiaorong Feng, Michael T Ferdig, Roland A Cooper, Jianbing Mu, Dror I Baruch, Alan J Magill, and Xin-zhuan Su. Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. *Nature*, 418(6895):320–323, July 2002.

[94] R K Mehlotra, H Fujioka, P D Roepe, O Janneh, L M Ursos, V Jacobs-Lorena, D T McNamara, et al. Evolution of a unique Plasmodium falciparum chloroquine-resistance phenotype in association with pfcrt polymorphism in Papua New Guinea and South America. *Proceedings of the National Academy of Sciences of the United States of America*, 98(22):12689–12694, October 2001.

[95] Arjen M Dondorp, François Nosten, Poravuth Yi, Debashish Das, Aung Phae Phyo, Joel Tarning, Khin Maung Lwin, et al. Artemisinin Resistance in Plasmodium falciparumMalaria. *New England Journal of Medicine*, 361(5):455–467, July 2009.

[96] Elizabeth A Ashley, Mehul Dhorda, Rick M Fairhurst, Chanaki Amaratunga, Parath Lim, Seila Suon, Sokunthea Sreng, et al. Spread of Artemisinin Resistance in Plasmodium falciparumMalaria. *New England Journal of Medicine*, 371(5):411–423, July 2014.

[97] MalariaGEN Plasmodium falciparum Community Project. Genomic epidemiology of artemisinin resistant malaria. *eLife*, 5:e08714, March 2016.

[98] Leslie Roberts. Malaria wars. *Science (New York, N.Y.)*, 352(6284):398–402– 404–5, April 2016.

[99] World Health Organization. Q and A on artemisinin resistance, 2016. URL `http://www.who.int/malaria/media/artemisinin-resistance-qa/en/`.

[100] P C Sabeti. Positive Natural Selection in the Human Lineage. *Science (New York, N.Y.)*, 312(5780):1614–1620, June 2006.

[101] Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A map of recent positive selection in the human genome. *PLoS biology*, 4(3):e72, March 2006.

[102] F Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, November 1989.

[103] Anna Ferrer-Admetlla, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular biology and evolution*, 31(5):1275–1291, May 2014.

[104] Pardis C Sabeti, David E Reich, John M Higgins, Haninah Z P Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, October 2002.

[105] Joseph J Vitti, Sharon R Grossman, and Pardis C Sabeti. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, 47(1):97–120, November 2013.

[106] Nicholas J Croucher, Claire Chewapreecha, William P Hanage, Simon R Harris, Lesley McGee, et al. Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape. *Genome biology and evolution*, 6(7): 1589–1602, June 2014.

[107] Pleuni S Pennings, Sergey Kryazhimskiy, and John Wakeley. Loss and recovery of genetic diversity in adapting populations of HIV. *PLoS genetics*, 10(1):e1004000, January 2014.

[108] Alfred Amambua-Ngwa, Daniel J Park, Sarah K Volkman, Kayla G Barnes, Amy K Bei, Amanda K Lukens, Papa Sene, et al. SNP genotyping identifies new signatures of selection in a deep sample of West African Plasmodium falciparum malaria parasites. *Molecular Biology and Evolution*, 29(11):3249–3253, November 2012.

[109] Victor A Mobegi, Craig W Duffy, Alfred Amambua-Ngwa, Kovana M Loua, Eugene Laman, Davis C Nwakanma, Bronwyn MacInnis, et al. Genome-wide analysis of selection on the malaria parasite Plasmodium falciparum in West African populations of differing infection endemicity. *Molecular biology and evolution*, 31(6):1490–1499, June 2014.

[110] Hanif Samad, Francesc Coll, Mark D Preston, Harold Ocholla, Rick M Fairhurst, and Taane G Clark. Imputation-based population genetics analysis of Plasmodium falciparum malaria parasites. *PLoS genetics*, 11(4):e1005131, April 2015.

[111] Craig W Duffy, Samuel A Assefa, James Abugri, Nicholas Amoako, Seth Owusu-Agyei, Thomas Anyorigiya, Bronwyn MacInnis, et al. Comparison of genomic signatures of selection on Plasmodium falciparum between different regions of a country with high malaria endemicity. *BMC genomics*, 16:527, July 2015.

[112] Lyndal Henden, Stuart Lee, Ivo Mueller, Alyssa Barry, and Melanie Bahlo. Detecting Selection Signals In Plasmodium falciparum Using Identity-By-Descent Analysis. *bioRxiv*, page 088039, November 2016.

[113] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5): 491–498, May 2011.

[114] Magnus Manske, Olivo Miotto, Susana Campino, Sarah Auburn, Jacob Almagro-Garcia, Gareth Maslen, Jack O'Brien, et al. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature*, 487(7407):375–379, July 2012.

[115] P W Messer. SLiM: simulating evolution with selection and linkage. *Genetics*, 194 (4):1037–1037, 2013.

[116] A L Hughes and F Verra. Very large long-term effective population size in the virulent human malaria parasite Plasmodium falciparum. *Proceedings. Biological sciences*, 268(1478):1855–1860, September 2001.

[117] Selina E R Bopp, Micah J Manary, A Taylor Bright, Geoffrey L Johnston, Neekesh V Dharia, Fabio L Luna, Susan McCormack, David Plouffe, Case W McNamara,

John R Walker, David A Fidock, Eros Lazzerini Denchi, and Elizabeth A Winzeler. Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS genetics*, 9(2):e1003293, 2013.

[118] Stuart Lee. moimix: an R package for evaluating multiplicity of infection in malaria parasites, 2016. URL `https://github.com/bahlolab/moimix`.

[119] Daniel E Neafsey, Stephen F Schaffner, Sarah K Volkman, Daniel Park, Philip Montgomery, Danny A Milner, Amanda Lukens, et al. Genome-wide SNP genotyping highlights the role of natural selection in Plasmodium falciparumpopulation divergence. *Genome Biology*, 9(12):R171, 2008.

[120] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, August 2006.

[121] Python Software Foundation. The python language reference, 2016. URL `https://www.python.org/`.

[122] A Miles and N Harding. scikit-allel:v0.20.3, 2016. URL `https://github.com/cggh/scikit-allel`.

[123] Xuanyao Liu, Chakravarthi Kanduri, Jaana Oikkonen, Kai Karma, Pirre Raijas, Liisa Ukkola-Vuoti, Yik-Ying Teo, and Irma Järvelä. Detecting signatures of positive selection associated with musical aptitude in the human genome. *Scientific reports*, 6:21198, February 2016.

[124] G Nepusz and G Csárdi. The igraph software package for complex network research. *Complex Systems*, 2006.

[125] Daniel J Park, Amanda K Lukens, Daniel E Neafsey, Stephen F Schaffner, Hsiao-Han Chang, Clarissa Valim, Ulf Ribacke, et al. Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. *PNAS*, 109(32):13052–13057, August 2012.

[126] Miriam K Laufer, Phillip C Thesing, Nicole D Eddington, Rhoda Masonga, Fraction K Dzinjalamala, Shannon L Takala, Terrie E Taylor, and Christopher V Plowe. Return of chloroquine antimalarial efficacy in Malawi. *The New England journal of medicine*, 355(19):1959–1966, November 2006.

[127] Davis C Nwakanma, Craig W Duffy, Alfred Amambua-Ngwa, Eniyou C Oriero, Kalifa A Bojang, Margaret Pinder, Chris J Drakeley, Colin J Sutherland, et al. Changes in malaria parasite drug resistance in an endemic population over a 25-year period with resulting genomic evidence of selection. *The Journal of infectious diseases*, 209(7):1126–1135, April 2014.

[128] Olivo Miotto, Jacob Almagro-Garcia, Magnus Manske, Bronwyn Macinnis, Susana Campino, Kirk A Rockett, Chanaki Amaratunga, Pharath Lim, Seila Suon, et al. Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. *Nature genetics*, 45(6):648–655, June 2013.

[129] Frédéric Ariey, Benoit Witkowski, Chanaki Amaratunga, Johann Beghain, Anne-Claire Langlois, Nimol Khim, Saorin Kim, Valentine Duru, Christiane Bouchier, et al. A molecular marker of artemisinin- resistant Plasmodium falciparum malaria. *Nature*, 505(7481):50–55, April 2015.

[130] Richard J Maude, Chea Nguon, Po Ly, Tol Bunkea, Pengby Ngor, Sara E Canavati de la Torre, Nicholas J White, Arjen M Dondorp, Nicholas P J Day, Lisa J White, and Char Meng Chuor. Spatial and temporal epidemiology of clinical malaria in Cambodia 2004-2013. *Malaria journal*, 13:385, September 2014.

[131] Shannon Takala-Harrison, Christopher G Jacob, Cesar Arze, Michael P Cummings, Joana C Silva, Arjen M Dondorp, Mark M Fukuda, Tran Tinh Hien, Mayfong Mayxay, Harald Noedl, et al. Independent emergence of artemisinin resistance mutations among Plasmodium falciparum in Southeast Asia. *The Journal of infectious diseases*, 211(5):670–679, March 2015.

[132] Alfred Amambua-Ngwa, Bakary Danso, Archibald Worwui, Sukai Ceesay, Nwakanma Davies, David Jeffries, Umberto D'Alessandro, and David Conway. Exceptionally long-range haplotypes in Plasmodium falciparum chromosome 6 maintained in an endemic African population. *Malaria journal*, 15(1):515, October 2016.

[133] M Isabel Veiga, Satish K Dhingra, Philipp P Henrich, Judith Straimer, Nina Gnädig, Anne-Catrin Uhlemann, Rowena E Martin, Adele M Lehane, and David A Fidock. Globally prevalent PfMDR1 mutations modulate Plasmodium falciparum susceptibility to artemisinin-based combination therapies. *Nature communications*, 7:11553, May 2016.

[134] Lynette Isabella Ochola-Oyier, John Okombo, Njoroge Wagatua, Jacob Ochieng, Kevin K Tetteh, Greg Fegan, Philip Bejon, and Kevin Marsh. Comparison of allele frequencies of Plasmodium falciparum merozoite antigens in malaria infections sampled in different years in a Kenyan population. *Malaria journal*, 15(1):261, May 2016.

[135] Jun Ohashi, Yuji Suzuki, Izumi Naka, Hathairad Hananantachai, and Jintana Patarapotikul. Diversifying selection on the thrombospondin-related adhesive protein (TRAP) gene of Plasmodium falciparum in Thailand. *PloS one*, 9(2):e90522, 2014.

[136] T J Anderson, B Haubold, J T Williams, J G Estrada-Franco, L Richardson, R Mollinedo, M Bockarie, J Mokili, S Mharakurwa, N French, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. *Molecular biology and evolution*, 17(10):1467–1482, October 2000.

[137] Rachel F Daniels, Stephen F Schaffner, Edward A Wenger, Joshua L Proctor, Hsiao-Han Chang, Wesley Wong, Nicholas Baro, Daouda Ndiaye, Fatou Ba Fall, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22):7067–7072, June 2015.

[138] Jane M Carlton, John H Adams, Joana C Silva, Shelby L Bidwell, Hernan Lorenzi, Elisabet Caler, Jonathan Crabtree, Samuel V Angiuoli, Emilio F Merino, Paolo Amedeo, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature*, 455(7214):757–763, October 2008.

[139] E J Feil, E C Holmes, D E Bessen, M S Chan, N P Day, M C Enright, R Goldstein, D W Hood, A Kalia, C E Moore, J Zhou, and B G Spratt. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 98(1):182–187, January 2001.

[140] Anastasia L Wise, Lin Gyi, and Teri A Manolio. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *American journal of human genetics*, 92(5):643–647, May 2013.

[141] Online Mendelian Inheritance in Man OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). URL `https://www.omim.org/statistics/entry`.

[142] Claude Libert, Lien Dejager, and Iris Pinheiro. The X chromosome in immune functions: when a chromosome makes the difference. *Nature reviews. Immunology*, 10(8):594–604, August 2010.

[143] Lyndal Henden, Saskia Freytag, Zaid Afawi, Sara Baldassari, Samuel F Berkovic, Francesca Bisulli, Laura Canafoglia, et al. Identity by descent fine mapping of familial adult myoclonus epilepsy (FAME) to 2p11.2-2q11.2. *Human genetics*, July 2016.

[144] Andrew Taylor Bright, Micah J Manary, Ryan Tewhey, Eliana M Arango, Tina Wang, Nicholas J Schork, Stephanie K Yanow, and Elizabeth A Winzeler. A High Resolution Case Study of a Patient with Recurrent Plasmodium vivax Infections Shows That Relapses Were Caused by Meiotic Siblings. *PLoS Neglected Tropical Diseases*, 8(6):e2882, June 2014.

[145] Antoine Claessens, William L Hamilton, Mihir Kekre, Thomas D Otto, Adnan Faizullabhoy, Julian C Rayner, and Dominic Kwiatkowski. Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis. *PLoS genetics*, 10(12):e1004812, December 2014.

[146] Elizabeth V Fowler, Marina Chavchich, Nanhua Chen, Jennifer M Peters, Dennis E Kyle, Michelle L Gatton, and Qin Cheng. Physical Linkage to Drug Resistance Genes Results in Conservation of varGenes among West Pacific Plasmodium falciparumIsolates. *The Journal of Infectious Diseases*, 194(7):939–948, October 2006.

[147] S A Assefa. *De novo assembly of the var multi-gene family in clinical samples of Plasmodium falciparum*. PhD thesis, University of Cambridge, 2013.

[148] K Lange. Mathematical and statistical methods for genetic analysis. New York, 1997.

[149] Picard tools. Picard tools, 2016. URL `http://broadinstitute.github.io/picard/`.

[150] H Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org*, 2013.

[151] S Andrews. FastQC: a quality control tool for high throughput sequence data, 2010.

[152] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, April 2012.

[153] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

Appendix A

XIBD: software for inferring pairwise identity by descent on the X chromosome

The following is a publication related to Chapter 2.

Bioinformatics, 32(15), 2016, 2389–2391
doi: 10.1093/bioinformatics/btw124
Advance Access Publication Date: 21 March 2016
Applications Note

OXFORD

Genetics and population analysis

XIBD: software for inferring pairwise identity by descent on the X chromosome

Lyndal Henden[1,2,*], David Wakeham[3] and Melanie Bahlo[1,2,4]

[1]Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia, [2]Department of Medical Biology, [3]School of Physics and [4]School of Mathematics and Statistics, University of Melbourne, Melbourne, VIC, Australia

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on October 1, 2015; revised on February 4, 2016; accepted on March 1, 2016

Abstract

Summary: XIBD performs pairwise relatedness mapping on the X chromosome using dense single nucleotide polymorphism (SNP) data from either SNP chips or next generation sequencing data. It correctly accounts for the difference in chromosomal numbers between males and females and estimates global relatedness as well as regions of the genome that are identical by descent (IBD). XIBD also generates novel graphical summaries of all pairwise IBD tracts for a cohort making it very useful for disease locus mapping.
Availability and implementation: XIBD is written in R/Rcpp and executed from shell scripts that are freely available from http://bioinf.wehi.edu.au/software/XIBD along with accompanying reference datasets.
Contact: henden.l@wehi.edu.au
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic regions that have been inherited from a common ancestor are said to be identical by descent (IBD). Identification of IBD regions has proven useful in many applications, including discovery and quantification of unknown or misspecified familial relatedness (McPeek and Sun, 2000) and disease mapping where a region inherited in multiple affected individuals is indicative of a critical region containing disease susceptibility genes (Albrechtsen *et al.*, 2009).

Several methods have been proposed for inferring IBD, however few allow for analysis of the X chromosome. This is unfortunate as the X chromosome allows for inference of more distant relatedness than autosomes since fewer recombination events occur, resulting in IBD segments that are sustained over longer periods of time. Current IBD analysis of the X chromosome is based on identity by state sharing, which is non-probabilistic and does not necessarily imply IBD (Browning and Browning, 2013; Gusev *et al.*, 2009). Here we present XIBD; the only hidden Markov model (HMM) that infers IBD on the X chromosome in addition to the autosomes, where IBD is detectable between individuals with a recent common ancestor, within 25 generations, rather than more distant relatedness. Many IBD methodologies rely on large cohorts to estimate allele frequency data or to infer linkage disequilibrium structure. XIBD can be applied to as few as two samples with the option to use HapMap allele frequency data for 11 populations. Furthermore, it can be applied to either single nucleotide polymorphism (SNP) chip or next generation sequencing (NGS) data (exome or genome wide).

2 Methods

XIBD implements a first order continuous time HMM to infer IBD between pairs of individuals using unphased genotype data, where time is the genetic map distance in centimorgans (cM). While the model is continuous time, IBD is estimated at the genotyped positions only. The memoryless assumption of a Markov process is unlikely to hold due to recombination, however McPeek and Sun (2000) have shown it to be a good approximation and like Albrechtsen *et al.* (2009) and Epstein *et al.* (2000), we also make this assumption.

The hidden states in the Markov model are the number of alleles shared IBD between a pair of individuals, which depend on the genders of the individuals being compared. Assuming the pair are not inbred, if at least one individual is male, then 0 or 1 allele will be shared IBD and the state space is $Z = \{0, 1\}$. Alternatively, if both individuals are female then 0, 1 or 2 alleles will be shared IBD with $Z = \{0, 1, 2\}$.

The initial state probabilities of sharing 0, 1 or 2 alleles IBD are denoted ω_0, ω_1 and ω_2 respectively, where $\sum_{i=0}^{2} \omega_i = 1$. These probabilities can be calculated using identity coefficients if relationships are known. We use IdCoefs (Abney, 2009) for autosomes and have implemented the equivalent for the X chromosome. Individuals may be distantly related with unknown relationships, therefore these values must be calculated using an alternative approach. We use the method-of-moments approach described in Purcell *et al.* (2007) to estimate these probabilities. The estimated values are used in the analysis as they can be accurately calculated for known and unknown relationships and avoid misspecified pedigrees leading to incorrect global estimates.

Since there are two state spaces in the model, we require two transition probability matrices. These can be computed by solving Kolmogorov's forward (or backward) equation given the transition rate matrices (Supplementary information, Eqs. S1 and S2). The transition rate matrices require the number of meiosis m separating the pair of individuals, estimated as in Purcell *et al.* (2007), and the recombination rate θ estimated by Ott (1999).

The genotypic state space for an individual also depends on their gender. Let A and a denote the reference allele and alternative allele respectively. Since male X chromosomes are haploid, they cannot have heterozygous genotype calls. Therefore the male genotypic state space is $G = \{A, a\}$ while the female genotypic state space is $G = \{AA, Aa, aa\}$. A pair of genotypes makes up the observation for each marker in the model. Hence, the observation state space differs for each of the three pairwise combinations of genders and this difference leads to three sets of emission probabilities (Supplementary Information, Table S1). The emission probabilities are functions of the individuals' genders, the state space, the observed genotype pair and the population allele frequencies. The population allele frequencies are calculated from either a reference dataset such as HapMap (http://hapmap.ncbi.nlm.nih.gov/) or the input dataset itself. We include a genotyping error term into our calculation of the emission probabilities as implemented by Albrechtsen *et al.* (2009) (Supplementary Information, Tables S2 and S3) and accommodate missing data.

Dense marker datasets allow for better detection of small IBD tracts and hence more distant relatedness. However, the presence of LD can result in unwanted background sharing. To avoid this, we allow the user to select one of two models to accommodate for LD.

1. Like Purcell *et al.* (2007), model 1 assumes the markers are in linkage equilibrium, which may require thinning of datasets prior to use. However datasets with denser markers in LD can be used at the expense of false IBD segments being reported (Brown *et al.*, 2012).
2. Like Albrechtsen *et al.* (2009), LD is implicitly accounted for in model 2 using conditional emission probabilities (Supplementary Information, Tables S4–S6). Pairwise LD between markers is calculated using the squared correlation (R^2) of reference genotypes using PLINK (Purcell *et al.*, 2007).

Markers in high LD ($R^2 > 0.99$) and markers with low minor allele frequency (MAF < 0.01) are removed from the analysis.

Unlike Purcell *et al.* (2007) and Albrechtsen *et al.* (2009), reference datasets are provided with XIBD. These datasets are the combined HapMap Phase II and III genotypes and allele frequencies from build 19 (The International HapMap Consortium, 2003); allowing the user to choose between the 11 HapMap population. Furthermore, given a homogeneous population, we allow the user to calculate the necessary frequencies from the input dataset itself or to specify their own homogeneous reference dataset of matching population.

Global relatedness estimates (ω_0, ω_1, ω_2 and m) are reported for each pair of individuals analyzed, as well as inferred IBD tracts from the Viterbi algorithm and posterior probabilities from the forward–backward algorithm (Rabiner, 1989). IBD results are reported in spreadsheets. These can be cumbersome to investigate when many shared regions are inferred. However, XIBD also produces novel graphical summaries that allow the user to visualize shared regions in multiple individuals (Fig. 1).

Unlike other algorithms, XIBD does not require a large cohort for accurate IBD inference. In Shaw *et al.* (2015), XIBD was implemented on a single pair of male individuals with X-linked intellectual disability, to verify that a detected variation was contained within a small shared IBD tract on the X chromosome, thus leading to the conclusion that the variant was causal, rather than a technical artifact.

Results from simulation studies can be found in the Supplementary Information, Figures S1 and S2. We note that pseudo-autosomal regions are excluded from analysis. Finally,

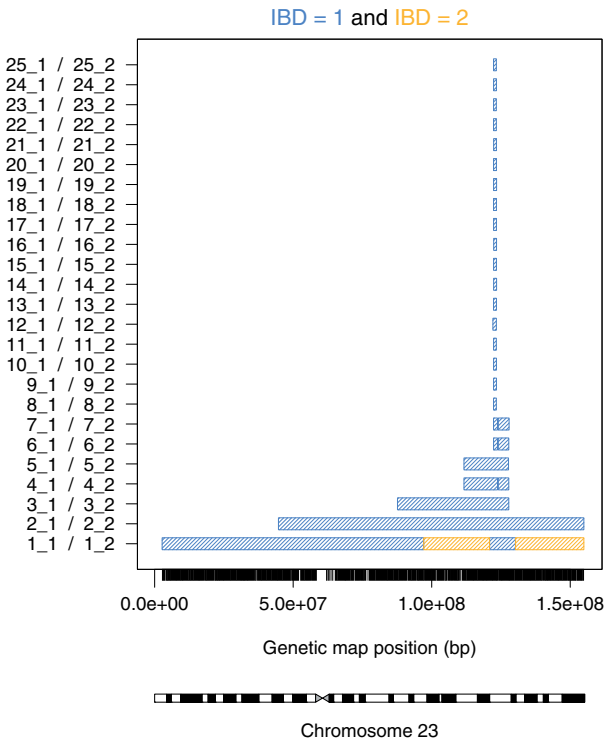


Fig. 1. Summary figure produced by XIBD of simulated IBD segments inherited up to 25 generations from the common ancestor. Pair identifiers are given on the *y*-axis. i.e. 23_1/23_2 are individuals 1 and 2 from generation 23, respectively. The *x*-axis displays the genetic map position in base-pairs with tick markers indicating the SNP positions. The ideogram for the X chromosome is below. Blue rectangles are regions where pairs share one allele IBD while yellow rectangles represent two alleles shared IBD

XIBD can also be extended for use on non-human haploid organisms to identify shared IBD tracts. We encourage feedback from users.

Funding

This work was supported by the Victorian Government's Operational Infrastructure Support Program and Australian Government NHMRC IRIIS [grant numbers 1002098 to M.B., 1054618 to M.B.]; The John and Patricia Farrant Scholarship [to L.H.]; and an Australian Postgraduate Award Scholarship [to L.H.].

Conflict of Interest: none declared.

References

Abney,M. (2009) A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics*, **25**, 1561–1563.

Albrechtsen,A. *et al*. (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol*., **33**, 266–274.

Brown,M.D. *et al*. (2012) Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*, **190**, 1447–1460.

Browning,B.L. and Browning,S.R. (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**, 459471.

Epstein,M.P. *et al*. (2000) Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet*., **67**, 1219–1231.

Gusev,A. *et al*. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res*., **12**, 318326.

McPeek,M.S. and Sun,L. (2000) Statistical test for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet*., **66**, 1076–1094.

Ott,J. (1999) Introduction and basic genetic principles. *Analysis of Human Genetic Linkage*. 3rd edn. Baltimore, London: Johns Hopkins University Press.

Purcell,S. *et al*. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Shaw,M. *et al*. (2015) Identical by descent L1CAM mutation in two apparently unrelated families with intellectual disability without L1 syndrome. *Eur. J. Med. Genet*., **58**, 364–368.

The International HapMap Consortium. (2003) The International HapMap Project. (2003) *Nature*, **426**, 789–796.

Appendix B

IBD coefficient calculation for the X chromosome

IBD coefficients can be calculated for autosomes given a pedigree as in Lange[148]. We extend this algorithm to the X chromosome below, excluding pseudoautosomal regions. The mathematical detail was performed by David Wakeham (member of the Bahlo lab in 2012/2013 summer). The corresponding R script was prepared by myself and is available as part of the XIBD R package.

The IBD coefficients, ω_0, ω_1 and ω_2, calculate the probability that two individuals will share 0, 1 or 2 alleles IBD at a randomly chosen locus, conditional on common ancestry. In order to calculate these probabilities, information on the specific allele configuration at a locus must be known, where allele configurations define *identity states*. Let FF, FM and MM denote a pair of females, a pair containing one female and one male, and a pair of males, respectively. The sets of identity states for each of the three combinations of pairs (by gender) are given in Figure B.1. When the maternal and paternal alleles are unknown (i.e. unphased data) the identity states can be reduced to form *condensed identity states* as in Figure B.2.

Let Δ_k denote the probability of a condensed identity state S_k. These probabilities are referred to as *identity coefficients* and are used to calculate the IBD coefficients. The IBD coefficients for two females (FF) expressed in terms of identity coefficients are

$$\begin{aligned}\omega_0 &= \Delta_2 + \Delta_4 + \Delta_6 + \Delta_9 \\ \omega_1 &= \Delta_3 + \Delta_5 + \Delta_8 \\ \omega_2 &= \Delta_1 + \Delta_7\end{aligned}$$

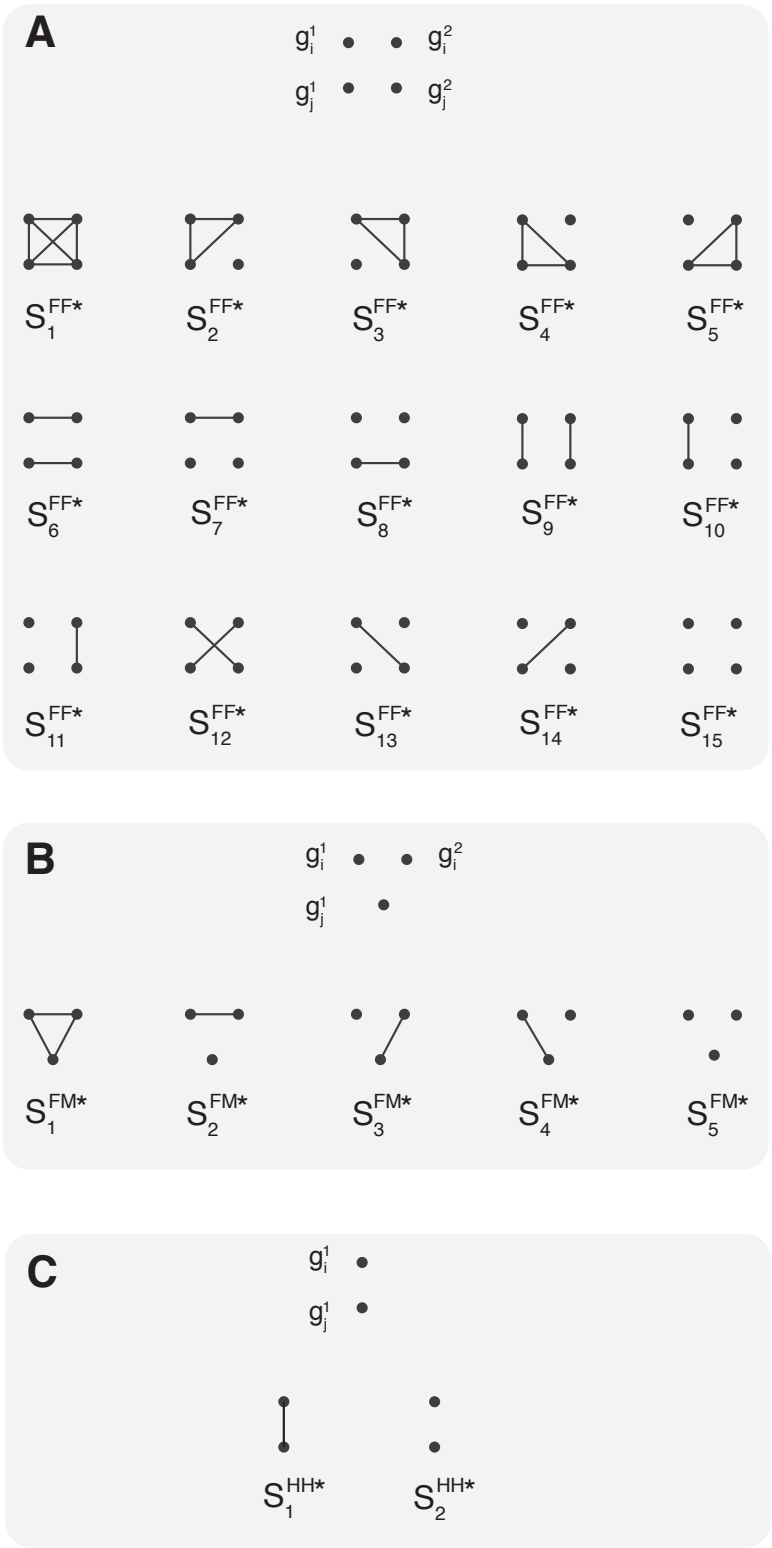


Figure B.1: Identity states for two individuals i and j with ordered genotypes g_i^1/g_i^2 and g_j^1/g_j^2. **A** The 15 identity states for pairs of female X chromosomes. **B** The 5 identity states for pairs of chromosomes where one chromosome belongs to a female and the other chromosome belongs to a male. **C** The 2 identity states for pairs of male chromosomes.

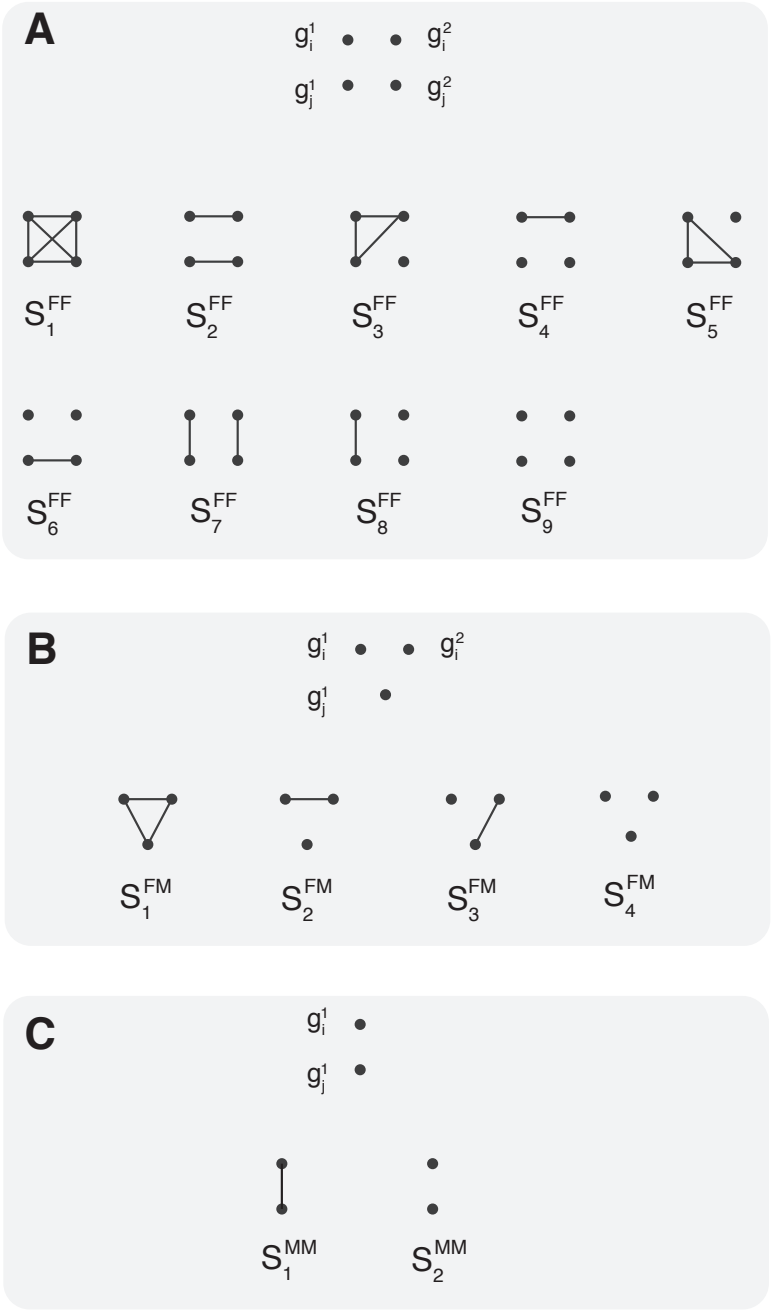


Figure B.2: Condensed identity states for unphased data. **A** The 9 condensed identity states for pairs of female chromosomes. **B** The 4 condensed identity states for pairs of chromosomes where one chromosome belongs to a female and the other chromosome belongs to a male. **C** The 2 condensed identity states for pairs of male chromosomes.

Similarly, for FM the IBD coefficients are

$$\begin{aligned}\omega_0 &= \Delta_2 + \Delta_4 \\ \omega_1 &= \Delta_1 + \Delta_3 \\ \omega_2 &= 0,\end{aligned}$$

and for MM, the relationship is simply;

$$\begin{aligned}\omega_0 &= \Delta_2 \\ \omega_1 &= \Delta_1 \\ \omega_2 &= 0.\end{aligned}$$

The relationship between IBD coefficients and identity coefficients is straightforward, however the calculation of identity coefficients is more complex. Let Ψ_k be the probability of a random condensed identity state, whereby Ψ_k describes a random sample with replacement of c_i alleles for both individuals at a (randomly chosen) locus on the X chromosome, where c_i is the number of copies of the X chromosome that individual i has. The probabilities Ψ_k can be expressed in terms of the identity coefficients Δ_k for two females as follows, where $\bar{\Psi}$ and $\bar{\Delta}$ are simply vectors containing Ψ_k and Δ_k, respectively.

$$\bar{\Psi}^{FF} = \begin{bmatrix} 1 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{8} & \frac{1}{16} & 0 \\ 0 & 1 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & \frac{1}{2} & \frac{1}{8} & \frac{1}{16} & \frac{1}{4} \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{8} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{8} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{8} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{8} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} \end{bmatrix} \bar{\Delta}^{FF}$$

The coefficient matrix is invertible (upper triangular, determinant product of diagonal elements), so we can express $\bar{\Delta}$ in terms of $\bar{\Psi}$ as

$$\bar{\Delta}^{FF} = \begin{bmatrix} 1 & 0 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & 1 & -\frac{1}{2} & -1 & -\frac{1}{2} & -1 & \frac{1}{2} & \frac{3}{4} & 1 \\ 0 & 0 & 2 & 0 & 0 & 0 & -2 & -1 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & 2 & 0 & -2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix} \bar{\Psi}^{FF}$$

We can do the same for a FM pair

$$\bar{\Psi}^{FM} = \begin{bmatrix} 1 & 0 & \frac{1}{4} & 0 \\ 0 & 1 & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{bmatrix} \bar{\Delta}^{FM} \quad \Rightarrow \quad \bar{\Delta}^{FM} = \begin{bmatrix} 1 & 0 & -\frac{1}{2} & 0 \\ 0 & 1 & -\frac{1}{2} & -1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \bar{\Psi}^{FM}$$

while the condition for MM is trivial, $\bar{\Psi}^{MM} = \bar{\Delta}^{MM}$. It just remains to calculate Ψ_k, which are done using *generalized kinship coefficients*. Generalized kinship coefficients are the probabilities of particular partitions of genes[148]. If P is a partition of sampled genes, we define $\Phi(P)$ to be the probability that the IBD relation induces P on the sampled genes. Let G_i^n denote the n-th allele sampled from individual i at a locus on chromosome X. For FF, i and j are both female and we sample G_i^1, G_i^2, G_j^1 and G_j^2, such that;

$$\begin{aligned}
\Psi_1^{FF} &= \phi(\{G_i^1, G_i^2, G_j^1, G_j^2\}) \\
\Psi_2^{FF} &= \phi(\{G_i^1, G_i^2\}, \{G_j^1, G_j^2\}) \\
\Psi_3^{FF} &= 2\phi(\{G_i^1, G_i^2, G_j^1\}, \{G_j^2\}) \\
\Psi_4^{FF} &= \phi(\{G_i^1, G_i^2\}, \{G_j^1\}, \{G_j^2\}) \\
\Psi_5^{FF} &= 2\phi(\{G_i^1, G_j^1, G_j^2\}, \{G_i^2\}) \\
\Psi_6^{FF} &= \phi(\{G_i^1\}, \{G_i^2\}, \{G_j^1, G_j^2\}) \\
\Psi_7^{FF} &= 2\phi(\{G_i^1, G_j^1\}, \{G_i^2, G_j^2\}) \\
\Psi_8^{FF} &= \phi(\{G_i^1, G_j^1\}, \{G_i^2\}, \{G_j^2\}) \\
\Psi_9^{FF} &= \phi(\{G_i^1\}, \{G_i^2\}, \{G_j^1\}, \{G_j^2\}).
\end{aligned}$$

For FM, let i be a female and j a male. Here the sampled genes are G_i^1, G_i^2 and G_j^1 and

$$
\begin{aligned}
\Psi_1^{FM} &= \phi(\{G_i^1, G_i^2, G_j^1\}) \\
\Psi_2^{FM} &= \phi(\{G_i^1, G_i^2\}, \{G_j^1\}) \\
\Psi_3^{FM} &= 2\phi(\{G_i^1, G_j^1\}, \{G_i^2\}) \\
\Psi_4^{FM} &= \phi(\{G_i^1\}, \{G_i^2\}, \{G_j^1\}).
\end{aligned}
$$

Finally, for MM we get

$$
\begin{aligned}
\Psi_1^{MM} &= \phi(\{G_i^1, G_j^1\}) \\
\Psi_2^{MM} &= \phi(\{G_i^1\}, \{G_j^1\}).
\end{aligned}
$$

We can determine $\Phi(P)$ using boundary conditions and recurrence relations, where we define a *block* to be a member of a partition; F, M to be female and male; and j and k to be the mother and father of i, respectively.

Boundary conditions

(B1) If F is involved in ≥ 3 or M is involved in ≥ 2 blocks, $\phi(P) = 0$. Cannot have that many X chromosome genes pairwise non-IBD.

(B2) If genes sampled from distinct founders occur in the same block, then $\phi(P) = 0$. Founder are not related.

(B3) If only founders contribute sampled genes, and neither (B1) or (B2) apply, then

$$\phi(P) = 2^{m2-m1}$$

where $m1$ is the total number of genes sampled from F founders and $m2$ is the number of F founders sampled. There are $m1 - m2$ comparison events, iid Bn(0.5)

Recurrence rules

(R1) Assume $G_i^1, \ldots, G_i^s$ are sampled from i for $s \geq 1$ and occur in one block, $P = \{\{G_i^1, \ldots, G_i^s, \ldots\}\} \cup P'$. If i is M, then

$$\phi(\{G_i^1, \ldots, G_i^s, \ldots\}, P') = \phi(\{G_j, \ldots\}, P')$$

since i receives their X gene from j. If i is F then

$$\begin{aligned}\phi(\{G_i^1,\dots,G_i^s,\dots\},P') &= (1-2^{1-s})\phi(\{G_j,G_k,\dots\},P')\\ &\quad +2^{-s}\phi(\{G_j,\dots\},P')\\ &\quad +2^{-s}\phi(\{G_k,\dots\},P')\end{aligned}$$

since there is 2^{-s} chance all sampled from j and 2^{-s} chance all sampled from k.

(R2) Now assume genes $G_i^1,\dots,G_i^s,G_i^{s+1},\dots,G_i^{s+t}$ sampled from F and $P=\{\{G_i^1,\dots,G_i^s,\dots\},\{G_i^{s+1},\dots,G_i^{s+t}\}\}\cup P'$. Then

$$\begin{aligned}\phi(\{G_i^1,\dots,G_i^s,\dots\},\{G_i^{s+1},\dots,G_j^{s+t}\},P') &= 2^{-(s+t)}\phi(\{G_j,\dots\}\{G_k,\dots\},P')\\ &\quad +2^{-(s+t)}\phi(\{G_k,\dots\}\{G_j,\dots\},P')\end{aligned}$$

This is because there is $2^{-(s+t)}$ chance that genes in the first block all come from j, and all genes in the second block come from k. Similarly with j and k swapped. There is no male analogue due to (B1).

Appendix C

Supplementary material for Chapter 6

C.1 Additional methods

C.1.1 Processing of Papua New Guinea Dataset

Picard Tools version 2.2.1 was used with the MarkIlluminaAdapters module to soft clip reads containing adapter sequences[149]. Following this paired-end reads were mapped to the Pf3D7 v3 reference genome with bwa-mem with the mark secondary hits option enabled[150]. The Genome Analysis Toolkit (GATK) version 3.5 was used with the RealignIndels walker to perform local realignment around intervals[113] and Picard-tools MarkDuplicates module was used to remove PCR duplicates. Next, GATK's BaseRecalibrator walker was used to correct base-quality scores using the entire *P. falciparum* genetic crosses version 1.0 data as known sites of variation[89]. Quality assessment of the aligned BAM files were performed with FastQC version 0.10.1[151] and Picard Tools CollectAlignmentMetrics and CollectInsertSizeMetrics module. Finally, coverage analysis was performed using GATK's Depth of Coverage walker with the conditions that reads had to have a mapping quality score of at least 20 and bases had to have a minimum quality of 20. As a result, 29 isolates were removed as less than 90% of their bases were not covered to at least 5 reads or did not map at all to the reference genome.

Variants were called using GATK's HaplotypeCaller walker in gVCF mode and genotypes were jointly called using GATK's GenotypeGVCF walker. Following this, variant quality score recalibration was performed for SNPs using the VariantRecalibrator walker.

The *P. falciparum genetic* crosses data was used as training data and the calibration model was trained using the QD, MQ, FS, SOR and DP tags in the VCF file. Annotation was performed using snpEff version 4.1 and a custom annotation was added to the final VCF file using the RegionType annotations obtained from the *P. falciparum* genetic crosses data using bcftools version 1.1 [152,153]. We applied the same filtering procedure to the isolates and SNP calls as for the MalariaGEN pf3k field isolates. The final VCF file consisted of 38 isolates and 29,631 SNPs.

C.2 Supplementary tables and figures

Appendix C Table 1. The number of isolates and SNPs before and after filtering procedures for the *P. falciparum* genetic cross dataset.

	Pre VCF filtering		Post VCF filtering		Post isoRelate filtering	
Cross	**No. isolates**	**No. SNPs**	**No. isolates**	**No. SNPs**	**No. isolates**	**No. SNPs**
3D7 x HB3	21	15,398	21	11,612	21	11,612
7G8 x GB4	40	14,426	40	10,903	40	10,903
HB3 x Dd2	37	14,914	37	10,637	37	10,637

Appendix C Table 2. The sample collection years and the number of isolates and SNPs after filtering procedures within each country.

				Post VCF filtering		Post isoRelate filtering	
Region	**Country**	**Collection year (min)[a]**	**Collection Year (max)[a]**	**No. isolates**	**No. SNPs**	**No. isolates**	**No. SNPs**
Africa	DR of the Congo	2013	2013	104	60,969	104	31,676
Africa	Ghana	2009	2013	563	258,289	563	28,483
Africa	Guinea	2011	2011	100	101,425	100	44,528
Africa	Malawi	2011	2011	357	154,265	357	40,225
Africa	Mali	2007	2007	84	46,668	84	19,339
Africa	Senegal	2001	2011	131	52,664	131	26,757
Africa	The Gambia	2008	2008	57	46,576	57	43,360
Southeast Asia	Bangladesh	2012	2012	45	34,708	45	32,322
Southeast Asia	Cambodia	2009	2012	521	58,987	521	28,448
Southeast Asia	Laos	2011	2012	84	46,447	84	33,006
Southeast Asia	Myanmar	2011	2013	57	33,179	57	29,997
Southeast Asia	Thailand	2011	2013	140	40,502	140	28,218
Southeast Asia	Vietnam	2011	2012	96	42,011	96	29,617
Oceania	PNG[b]	2007	2007	38	29,631	37	18,270

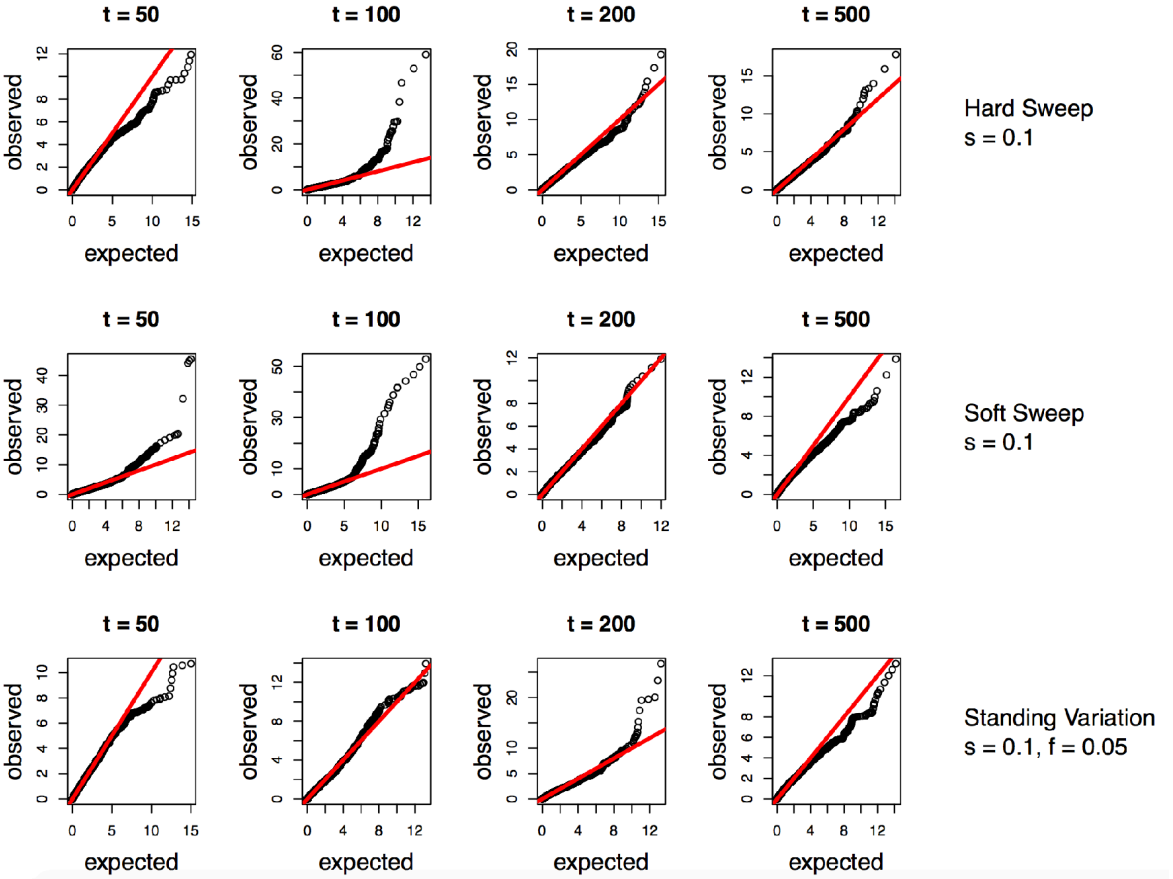
[a] Minimum and maximum collection years were obtained from pf3k metadata.
[b] Papua New Guinea is represented by the acronym PNG.

Appendix C Table 3. The number of isolates and SNPs included in the IBD analyses between pairs of countries.

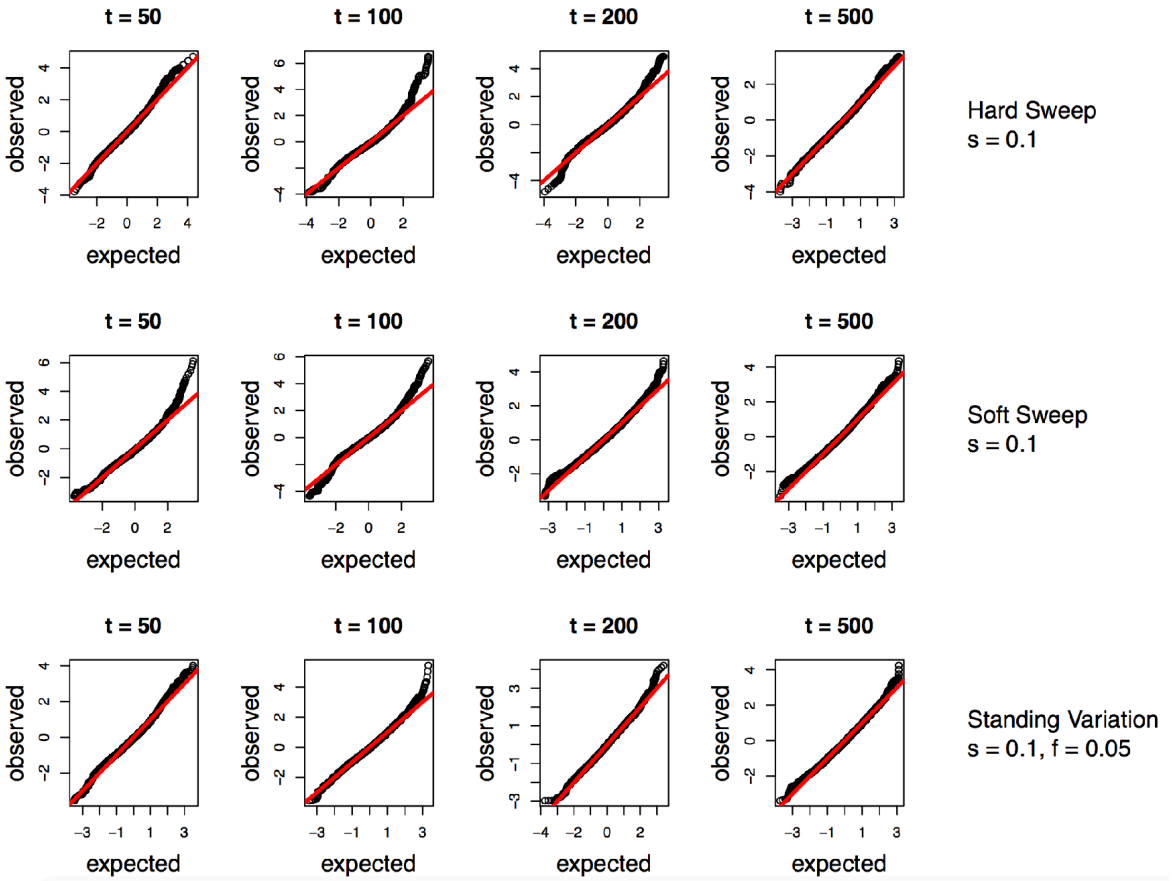
Region A	Region B	Country A	Country B	No. isolates	No. SNPs
Africa	Africa	DR of the Congo	Ghana	667	19,167
Africa	Africa	DR of the Congo	Guinea	204	19,236
Africa	Africa	DR of the Congo	Malawi	461	20,110
Africa	Africa	DR of the Congo	Mali	188	16,439
Africa	Africa	DR of the Congo	Senegal	235	16,529
Africa	Africa	DR of the Congo	The Gambia	161	16,192
Africa	Southeast Asia	DR of the Congo	Bangladesh	149	10,970
Africa	Southeast Asia	DR of the Congo	Cambodia	625	7,775
Africa	Southeast Asia	DR of the Congo	Laos	188	7,814
Africa	Southeast Asia	DR of the Congo	Myanmar	161	7,211
Africa	Southeast Asia	DR of the Congo	Thailand	244	7,353
Africa	Southeast Asia	DR of the Congo	Vietnam	200	7,390
Africa	Oceania	DR of the Congo	PNG	142	2,813
Africa	Africa	Ghana	Guinea	663	26,656
Africa	Africa	Ghana	Malawi	920	26,595
Africa	Africa	Ghana	Mali	647	16,127
Africa	Africa	Ghana	Senegal	694	18,138
Africa	Africa	Ghana	The Gambia	620	19,478
Africa	Southeast Asia	Ghana	Bangladesh	608	11,689
Africa	Southeast Asia	Ghana	Cambodia	1084	12,457
Africa	Southeast Asia	Ghana	Laos	647	10,334
Africa	Southeast Asia	Ghana	Myanmar	620	9,398
Africa	Southeast Asia	Ghana	Thailand	703	10,680
Africa	Southeast Asia	Ghana	Vietnam	659	10,046
Africa	Oceania	Ghana	PNG	600	4,118
Africa	Africa	Guinea	Malawi	457	29,138
Africa	Africa	Guinea	Mali	184	18,079
Africa	Africa	Guinea	Senegal	231	18,497
Africa	Africa	Guinea	The Gambia	157	21,457
Africa	Southeast Asia	Guinea	Bangladesh	145	11,937
Africa	Southeast Asia	Guinea	Cambodia	621	10,270
Africa	Southeast Asia	Guinea	Laos	184	11,458
Africa	Southeast Asia	Guinea	Myanmar	157	10,129
Africa	Southeast Asia	Guinea	Thailand	240	10,428
Africa	Southeast Asia	Guinea	Vietnam	196	10,756
Africa	Oceania	Guinea	PNG	134	4,834
Africa	Africa	Malawi	Mali	441	14,493

Africa	Africa	Malawi	Senegal	488	16,373
Africa	Africa	Malawi	The Gambia	414	19,049
Africa	Southeast Asia	Malawi	Bangladesh	402	12,405
Africa	Southeast Asia	Malawi	Cambodia	878	12,233
Africa	Southeast Asia	Malawi	Laos	441	12,638
Africa	Southeast Asia	Malawi	Myanmar	414	10,837
Africa	Southeast Asia	Malawi	Thailand	497	11,657
Africa	Southeast Asia	Malawi	Vietnam	453	11,591
Africa	Oceania	Malawi	PNG	391	4,988
Africa	Africa	Mali	Senegal	215	17,458
Africa	Africa	Mali	The Gambia	141	13,899
Africa	Southeast Asia	Mali	Bangladesh	129	8,997
Africa	Southeast Asia	Mali	Cambodia	605	5,941
Africa	Southeast Asia	Mali	Laos	168	5,805
Africa	Southeast Asia	Mali	Myanmar	141	5,622
Africa	Southeast Asia	Mali	Thailand	224	5,507
Africa	Southeast Asia	Mali	Vietnam	180	5,555
Africa	Oceania	Mali	PNG	122	1,945
Africa	Africa	Senegal	The Gambia	188	16,364
Africa	Southeast Asia	Senegal	Bangladesh	176	9,275
Africa	Southeast Asia	Senegal	Cambodia	652	6,486
Africa	Southeast Asia	Senegal	Laos	215	6,315
Africa	Southeast Asia	Senegal	Myanmar	188	5,885
Africa	Southeast Asia	Senegal	Thailand	271	6,017
Africa	Southeast Asia	Senegal	Vietnam	227	6,011
Africa	Oceania	Senegal	PNG	169	2,274
Africa	Southeast Asia	The Gambia	Bangladesh	102	9,882
Africa	Southeast Asia	The Gambia	Cambodia	578	7,305
Africa	Southeast Asia	The Gambia	Laos	141	7,979
Africa	Southeast Asia	The Gambia	Myanmar	114	7,305
Africa	Southeast Asia	The Gambia	Thailand	197	7,398
Africa	Southeast Asia	The Gambia	Vietnam	153	7,504
Africa	Oceania	The Gambia	PNG	94	2,984
Southeast Asia	Southeast Asia	Bangladesh	Cambodia	566	13,736
Southeast Asia	Southeast Asia	Bangladesh	Laos	129	15,029
Southeast Asia	Southeast Asia	Bangladesh	Myanmar	102	14,751
Southeast Asia	Southeast Asia	Bangladesh	Thailand	185	14,666
Southeast Asia	Southeast Asia	Bangladesh	Vietnam	141	14,207
Southeast Asia	Oceania	Bangladesh	PNG	83	4,218
Southeast Asia	Southeast Asia	Cambodia	Laos	605	24,921
Southeast Asia	Southeast Asia	Cambodia	Myanmar	578	17,479

Southeast Asia	Southeast Asia	Cambodia	Thailand	661	21,199
Southeast Asia	Southeast Asia	Cambodia	Vietnam	617	23,394
Southeast Asia	Oceania	Cambodia	PNG	558	5,395
Southeast Asia	Southeast Asia	Laos	Myanmar	141	19,045
Southeast Asia	Southeast Asia	Laos	Thailand	224	21,164
Southeast Asia	Southeast Asia	Laos	Vietnam	180	26,419
Southeast Asia	Oceania	Laos	PNG	121	5,994
Southeast Asia	Southeast Asia	Myanmar	Thailand	197	21,293
Southeast Asia	Southeast Asia	Myanmar	Vietnam	153	18,666
Southeast Asia	Oceania	Myanmar	PNG	94	5,124
Southeast Asia	Southeast Asia	Thailand	Vietnam	236	21,069
Southeast Asia	Oceania	Thailand	PNG	177	5,572
Southeast Asia	Oceania	Vietnam	PNG	133	5,734



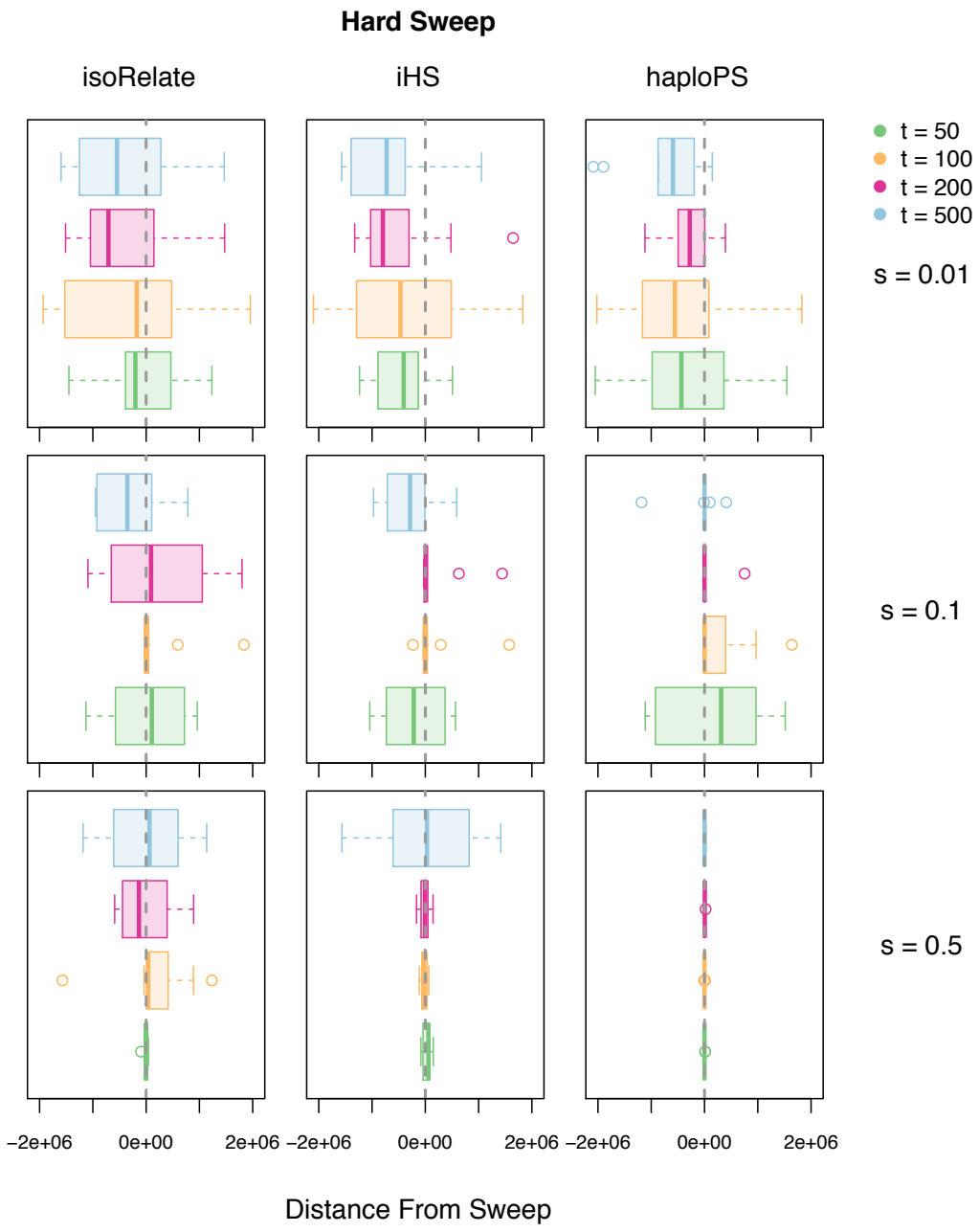
Appendix C Figure 1. Chi-squared QQ plots from isoRelate's selection statistic over three scenarios of positive selection. These plots correspond to one replicate of each scenario, where *t* is the number of generations since the sweep was introduced.



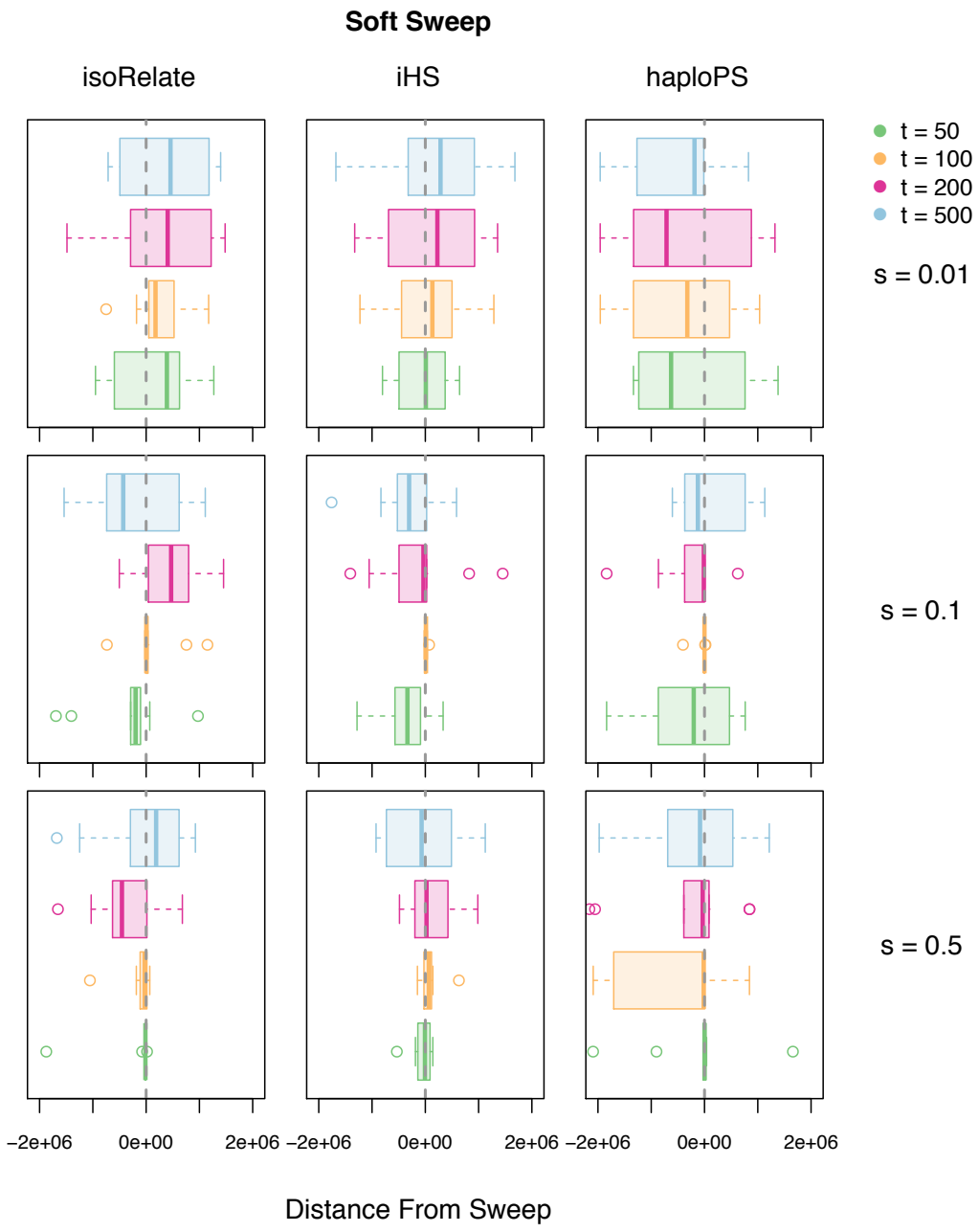
Appendix C Figure 2. Normal QQ plots from iHS's selection statistic over three scenarios of positive selection. These plots correspond to one replicate of each scenario, where *t* is the number of generations since the sweep was introduced. The same replicates were used in these figures as in Appendix C Figure 1.

Appendix C Table 4. The average percentage of isolates with various simulated MOI. The percentages were averaged over all 150 datasets corresponding to the parameter combinations assessed.

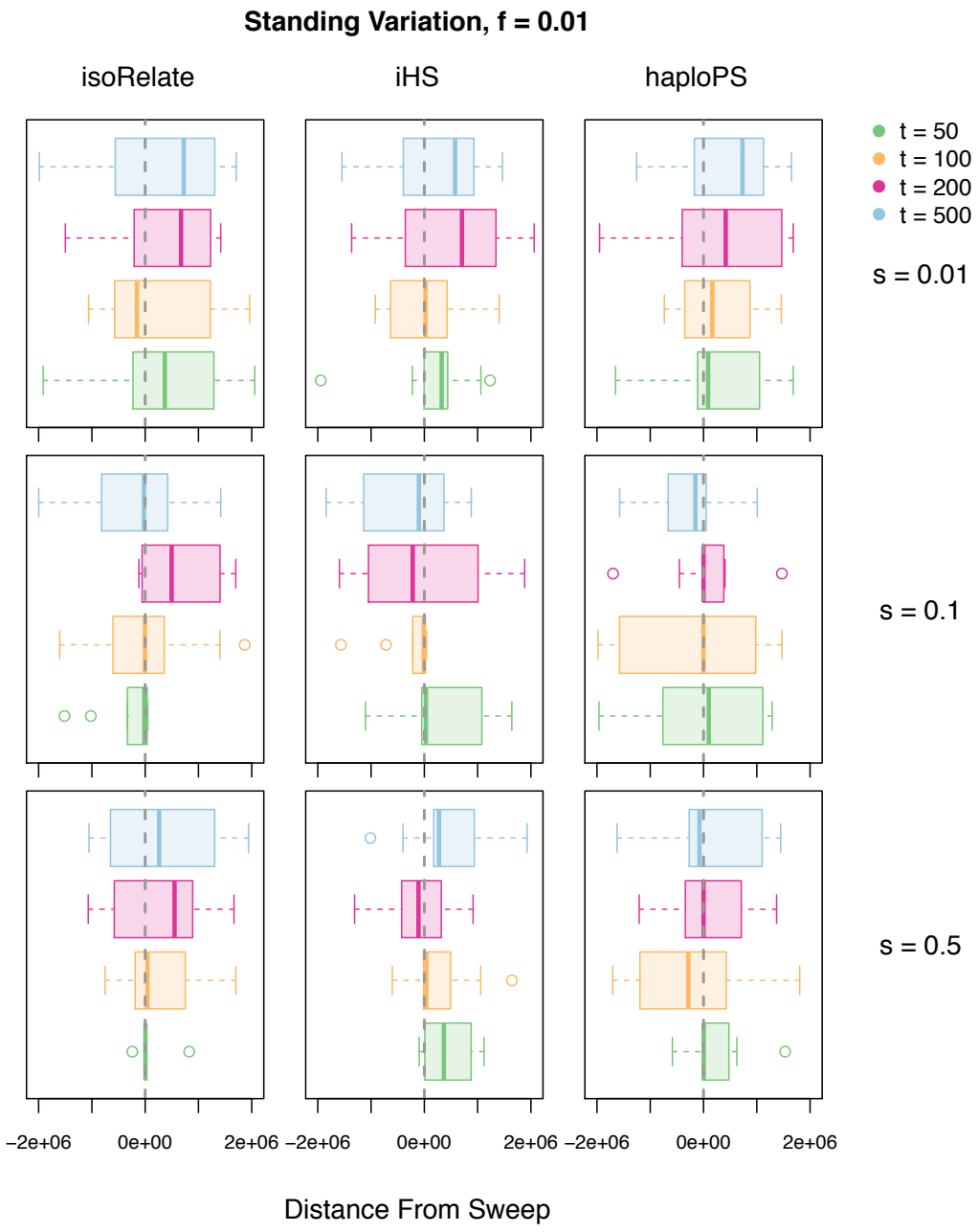
MOI						
1	**2**	**3**	**4**	**5**	**6**	**7**
56.81%	30.03%	10.01%	2.55%	0.48%	0.11%	0.01%



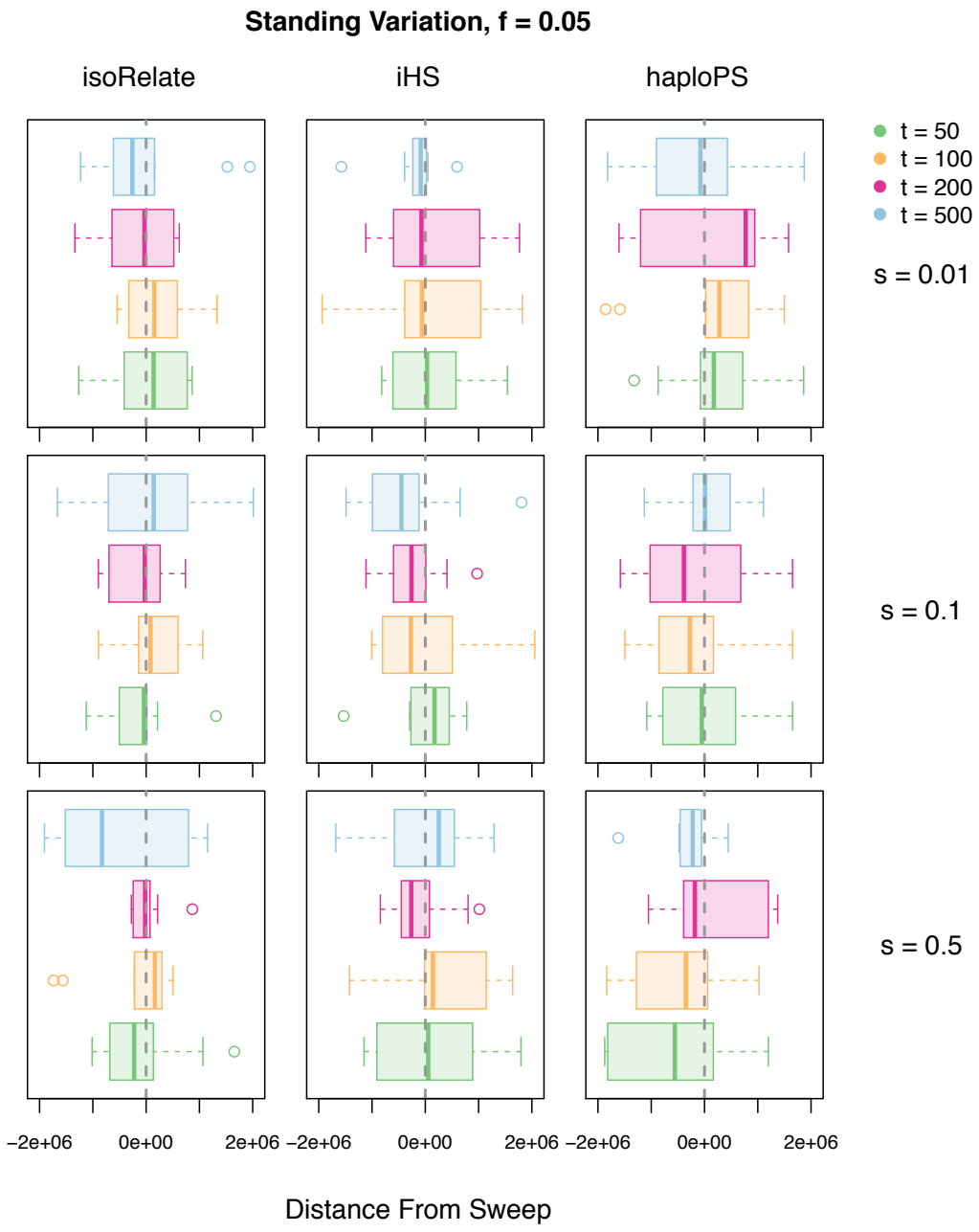
Appendix C Figure 3. Simulation results from hard sweeps for different selection coefficients when MOI ≥ 1 is incorporated into the simulation. isoRelate was run on all isolates while iHS and haploPS were run on MOI = 1 isolates only. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest –log10 p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario.



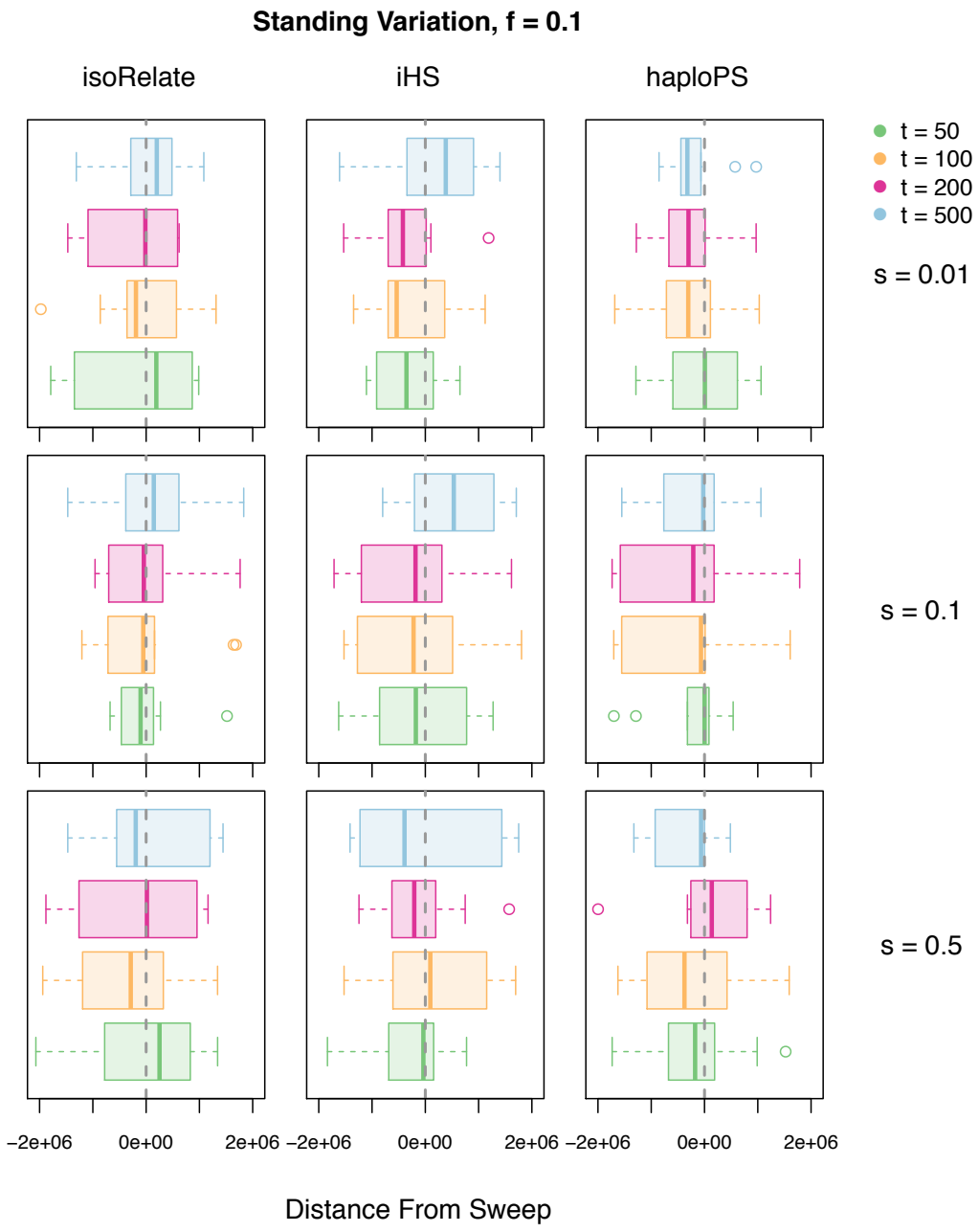
Appendix C Figure 4. Simulation results from soft sweeps for different selection coefficients when MOI ≥ 1 is incorporated into the simulation. isoRelate was run on all isolates while iHS and haploPS were run on MOI = 1 isolates only. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest –log10 p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario.



Appendix C Figure 5. Simulation results from standing variation with initial allele frequency f = 0.01 for different selection coefficients when MOI ≥ 1 is incorporated into the simulation. isoRelate was run on all isolates while iHS and haploPS were run on MOI = 1 isolates only. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest –log10 p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario.



Appendix C Figure 6. Simulation results from standing variation with initial allele frequency f = 0.05 for different selection coefficients when MOI ≥ 1 is incorporated into the simulation. isoRelate was run on all isolates while iHS and haploPS were run on MOI = 1 isolates only. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest –log10 p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario.



Appendix C Figure 7. Simulation results from standing variation with initial allele frequency f = 0.1 for different selection coefficients when MOI ≥ 1 is incorporated into the simulation. isoRelate was run on all isolates while iHS and haploPS were run on MOI = 1 isolates only. Boxplots show the distance between the genetic position of the sweep and the SNP with the largest –log10 p-value (isoRelate and iHS) or the SNP with most number of significant haplotypes overlapping it (haploPS), calculated across 10 replicates for each scenario.

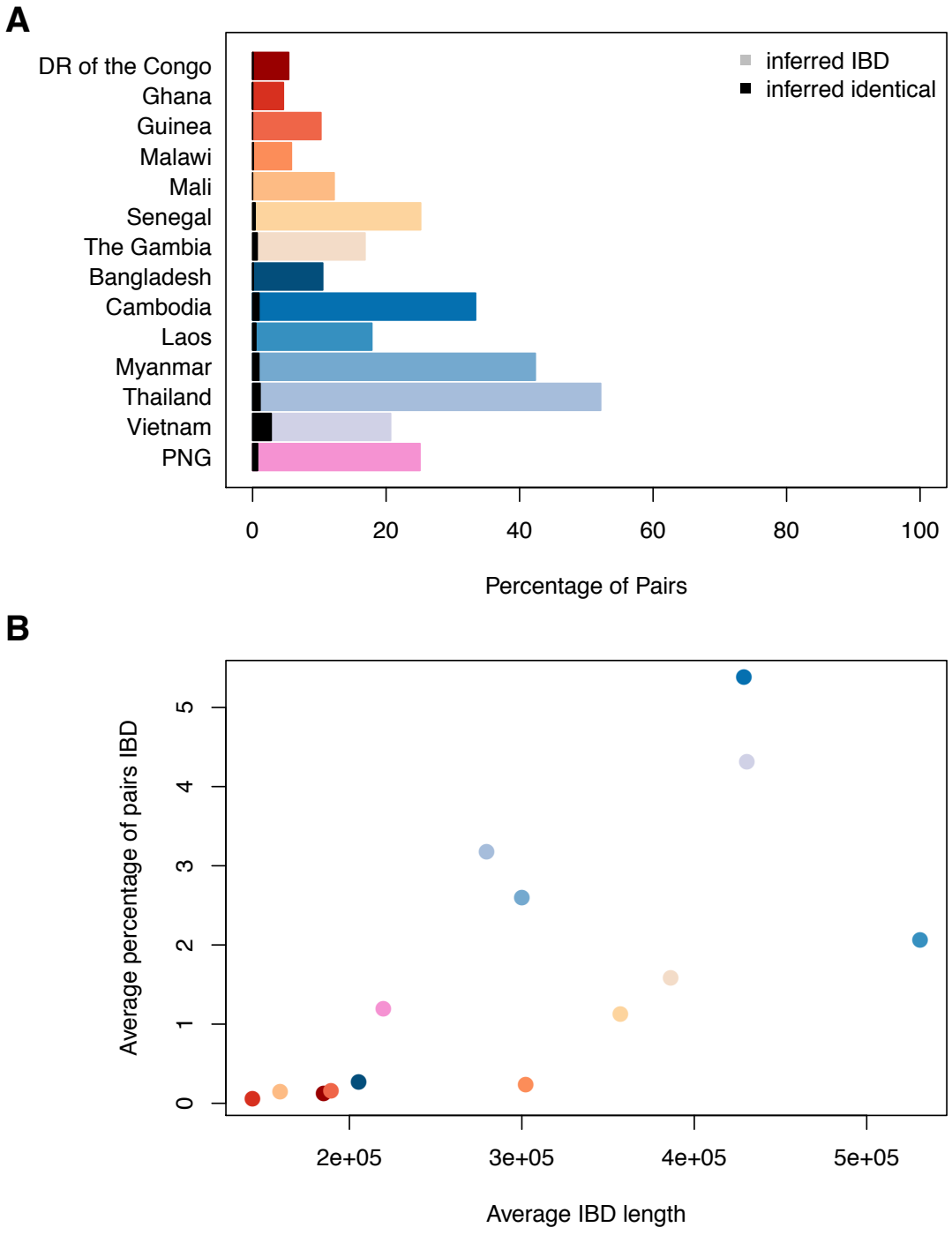
Appendix C Table 5. The number of isolates with MOI = 1 and MOI > 1 within each country.

Region	Country	MOI = 1	MO1 > 1
Africa	DR of the Congo	47	57
Africa	Ghana	243	320
Africa	Guinea	51	49
Africa	Malawi	134	223
Africa	Mali	40	44
Africa	Senegal	110	21
Africa	The Gambia	40	17
Southeast Asia	Bangladesh	21	24
Southeast Asia	Cambodia	396	125
Southeast Asia	Laos	49	35
Southeast Asia	Myanmar	45	12
Southeast Asia	Thailand	107	33
Southeast Asia	Vietnam	69	27
Oceania	PNG	31	7

Appendix C Table 6. Summary of relatedness between pairs of isolates within the same country.

Region	Country	No. isolates	No. pairs	% of pairs IBD[a]	% of pairs identical[b]	Ave. % of pairs IBD per SNP[c]	Ave. % of genome IBD[d]	Ave. length of IBD (kb)[e]
Africa	DR of the Congo	104	5,356	5.41	0.06	0.12	1.06	185
Africa	Ghana	563	158,203	4.62	0.01	0.06	0.78	144
Africa	Guinea	100	4,950	10.24	0	0.16	1.46	189
Africa	Malawi	357	63,546	5.82	0.11	0.24	2.24	302
Africa	Mali	84	3,486	12.22	0	0.15	0.83	160
Africa	Senegal	131	8,515	25.18	0.38	1.13	2.94	357
Africa	The Gambia	57	1,596	16.85	0.69	1.59	5.44	386
Southeast Asia	Bangladesh	45	990	10.51	0.1	0.27	1.3	205
Southeast Asia	Cambodia	521	135,460	33.41	0.95	5.38	13.72	429
Southeast Asia	Laos	84	3,486	17.87	0.49	2.06	8.86	531
Southeast Asia	Myanmar	57	1,596	42.36	0.94	2.6	3.82	300
Southeast Asia	Thailand	140	97,30	52.15	1.12	3.18	3.8	280
Southeast Asia	Vietnam	96	4,560	20.68	2.79	4.31	8.25	431
Oceania	PNG	37	666	25.08	0.75	1.19	1.68	220

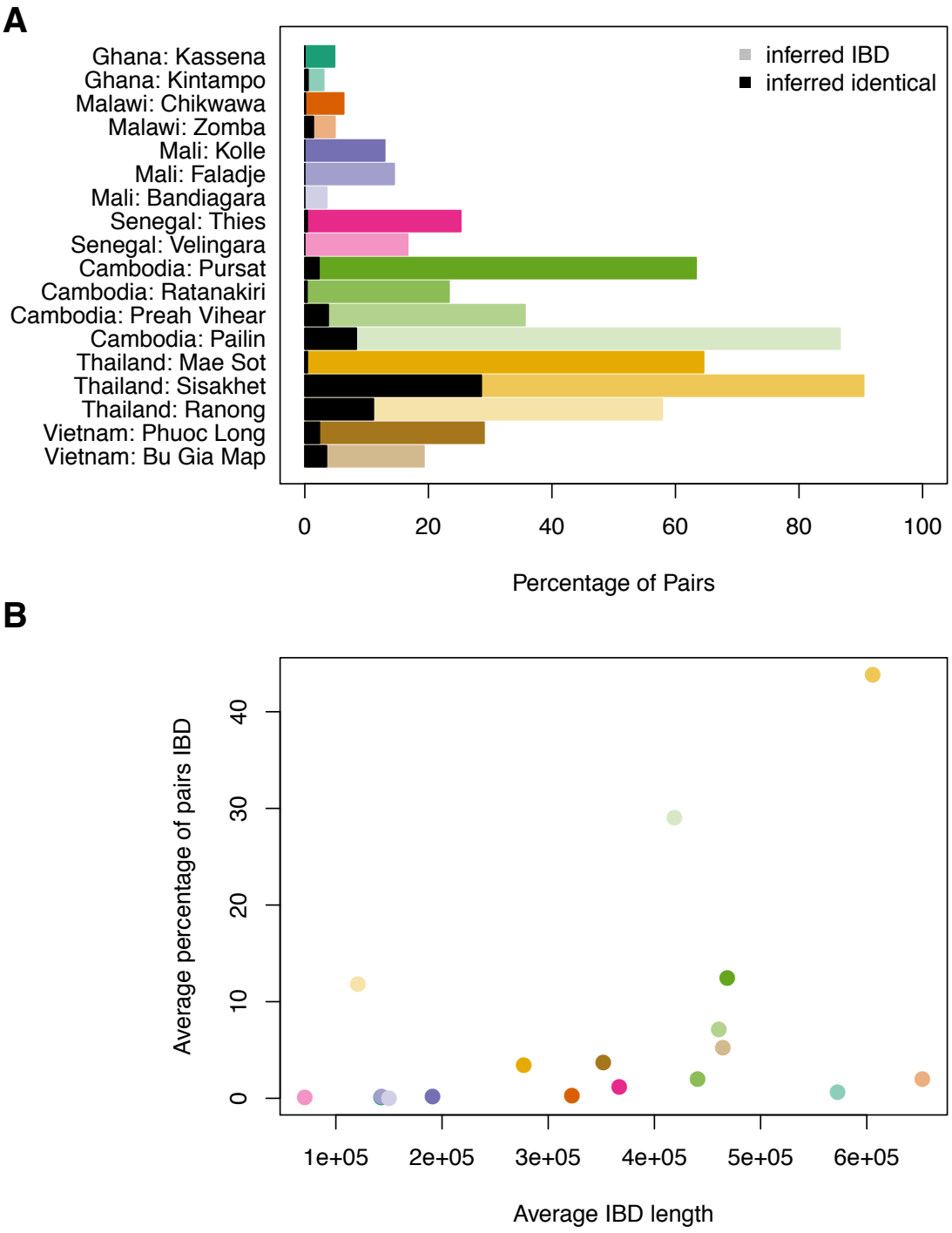
[a] Percentage of all pairs inferred IBD at any genomic location.
[b] Percentage of all pairs with identical genomes.
[c] Average percentage of pairs IBD calculated genome-wide.
[d] Average percentage of genome IBD calculated from IBD pairs only, excluding identical pairs.
[e] Average length of inferred IBD segments (kb), excluding segments from identical pairs.



Appendix C Figure 8. Summary of Appendix C Table 6. **A.** The percentage of pairs with any inferred IBD at least 50 kb in length within each country and the percentage of pairs with identical genomes. **B.** The average percentage of pairs IBD against the average IBD length in base-pairs.

Appendix C Table 7. Summary of relatedness between pairs of isolates within the same country, stratified by study location. Only countries with multiple study sites are displayed.

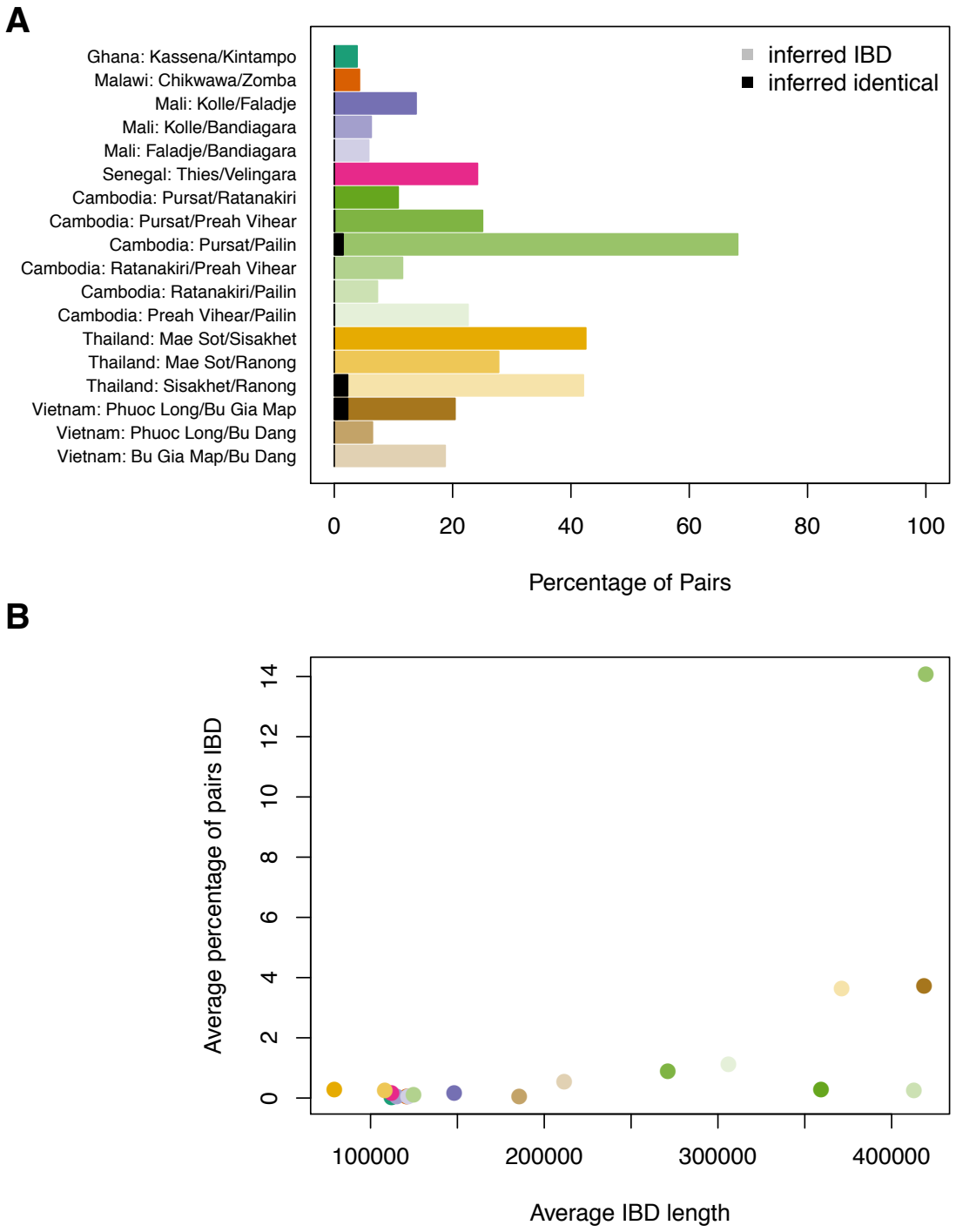
Region	Country	Site	No. isolates	No. pairs	% of pairs IBD	% of pairs identical	Ave. % of pairs IBD per SNP	Ave. % of genome IBD	Ave. length of IBD (kb)
Africa	Ghana	Kassena	501	125,250	4.83	0.01	0.06	0.78	143
Africa	Ghana	Kintampo	62	1,891	3.07	0.53	0.68	6.04	572
Africa	Malawi	Chikwawa	310	47,895	6.31	0.11	0.27	2.35	322
Africa	Malawi	Zomba	47	1,081	4.9	1.39	1.99	16.89	653
Africa	Mali	Kolle	46	1,035	12.95	0	0.18	1.09	191
Africa	Mali	Faladje	30	435	14.48	0	0.17	0.7	143
Africa	Mali	Bandiagara	8	28	3.57	0	0.03	0.71	150
Africa	Senegal	Thies	127	8,001	25.25	0.4	1.19	3.09	367
Africa	Senegal	Velingara	4	6	16.67	0	0.06	0.34	71
Southeast Asia	Cambodia	Pursat	219	23,871	63.35	2.3	12.48	16.54	469
Southeast Asia	Cambodia	Ratanakiri	134	8,911	23.36	0.36	1.97	6.74	441
Southeast Asia	Cambodia	Preah Vihear	86	3,655	35.65	3.8	7.14	10.3	461
Southeast Asia	Cambodia	Pailin	82	3,321	86.66	8.34	29.08	27.18	419
Southeast Asia	Thailand	Mae Sot	100	4,950	64.57	0.42	3.46	4.48	277
Southeast Asia	Thailand	Sisakhet	21	210	90.48	28.57	43.84	25.06	606
Southeast Asia	Thailand	Ranong	19	171	57.89	11.11	11.83	1.17	121
Southeast Asia	Vietnam	Phuoc Long	31	465	29.03	2.37	3.71	4.74	352
Southeast Asia	Vietnam	Bu Gia Map	64	2,016	19.3	3.52	5.22	10.5	464



Appendix C Figure 9. Summary of Appendix C Table 7. **A.** The percentage of pairs with any inferred IBD at least 50 kb in length within each site and the percentage of pairs with identical genomes. **B.** The average percentage of pairs IBD against the average IBD length in base-pairs.

Appendix C Table 8. Summary of relatedness between pairs of isolates from different sites within a country.

Region	Country	Site A	Site B	No. isolates	No. pairs	% of pairs IBD	% of pairs identical	Ave. % of pairs IBD per SNP	Ave. % of genome IBD	Ave. length of IBD (kb)
Africa	Ghana	Kassena	Kintampo	563	31,062	3.86	0	0.03	0.55	112
Africa	Malawi	Chikwawa	Zomba	357	14,570	4.26	0	0.04	0.84	121
Africa	Mali	Kolle	Faladje	76	1,380	13.84	0	0.16	0.74	148
Africa	Mali	Kolle	Bandiagara	54	368	6.25	0	0.05	0.57	115
Africa	Mali	Faladje	Bandiagara	38	240	5.83	0	0.04	0.58	122
Africa	Senegal	Thies	Velingara	131	508	24.21	0	0.17	0.6	112
Southeast Asia	Cambodia	Pursat	Ratanakiri	353	29,346	10.81	0	0.29	2.52	359
Southeast Asia	Cambodia	Pursat	Preah Vihear	305	18,834	25.07	0.07	0.89	3.04	271
Southeast Asia	Cambodia	Pursat	Pailin	301	17,958	68.2	1.49	14.08	19.41	420
Southeast Asia	Cambodia	Ratanakiri	Preah Vihear	220	11,524	11.54	0	0.11	0.71	125
Southeast Asia	Cambodia	Ratanakiri	Pailin	216	10,988	7.3	0	0.25	3.35	413
Southeast Asia	Cambodia	Preah Vihear	Pailin	168	7,052	22.62	0.04	1.12	4.44	306
Southeast Asia	Thailand	Mae Sot	Sisakhet	121	2,100	42.52	0	0.29	0.48	79
Southeast Asia	Thailand	Mae Sot	Ranong	119	1,900	27.79	0	0.26	0.65	108
Southeast Asia	Thailand	Sisakhet	Ranong	40	399	42.11	2.26	3.65	3.43	371
Southeast Asia	Vietnam	Phuoc Long	Bu Gia Map	95	1,984	20.41	2.27	3.72	7.7	419
Southeast Asia	Vietnam	Phuoc Long	Bu Dang	32	31	6.45	0	0.06	0.88	186
Southeast Asia	Vietnam	Bu Gia Map	Bu Dang	65	64	18.75	0	0.54	2.59	212



Appendix C Figure 10. Summary of Appendix C Table 8. **A.** The percentage of pairs with any inferred IBD at least 50 kb in length between sites within a country and the percentage of pairs with identical genomes. **B.** The average percentage of pairs IBD against the average IBD length in base-pairs.

Appendix C Table 9. Summary of relatedness between pairs of isolates from different countries.

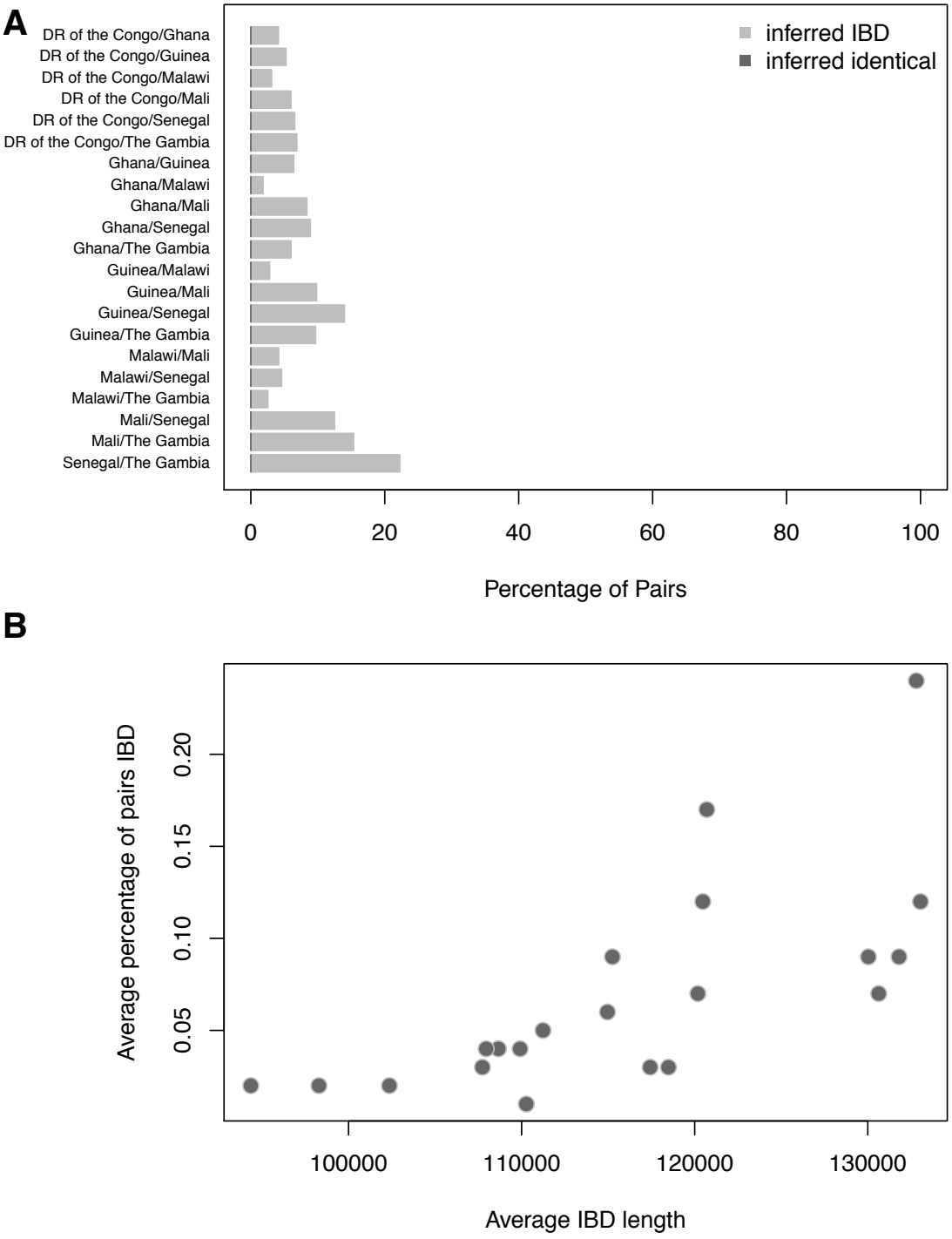
Region A	Region B	Country A	Country B	No. isolates	No. pairs	% of pairs IBD	% of pairs identical	Ave. % of pairs IBD per SNP	Ave. % of genome IBD	Ave. length of IBD (kb)
Africa	Africa	DR of the Congo	Ghana	667	58,552	4.12	0	0.03	0.52	108
Africa	Africa	DR of the Congo	Guinea	204	10,400	5.25	0	0.04	0.52	109
Africa	Africa	DR of the Congo	Malawi	461	37,128	3.1	0	0.02	0.49	102
Africa	Africa	DR of the Congo	Mali	188	8,736	5.99	0	0.04	0.52	108
Africa	Africa	DR of the Congo	Senegal	235	13,624	6.55	0	0.04	0.54	110
Africa	Africa	DR of the Congo	The Gambia	161	5,928	6.87	0	0.05	0.53	111
Africa	Africa	Ghana	Guinea	663	56,300	6.41	0	0.07	0.79	131
Africa	Africa	Ghana	Malawi	920	200,991	1.83	0	0.01	0.58	110
Africa	Africa	Ghana	Mali	647	47,292	8.37	0	0.09	0.64	130
Africa	Africa	Ghana	Senegal	694	73,753	8.88	0	0.07	0.62	120
Africa	Africa	Ghana	The Gambia	620	32,091	6.01	0	0.06	0.57	115
Africa	Africa	Guinea	Malawi	457	35,700	2.81	0	0.02	0.63	98
Africa	Africa	Guinea	Mali	184	8,400	9.83	0	0.09	0.65	132
Africa	Africa	Guinea	Senegal	231	13,100	13.98	0	0.12	0.65	120
Africa	Africa	Guinea	The Gambia	157	5,700	9.67	0	0.09	0.6	115
Africa	Africa	Malawi	Mali	441	29,988	4.16	0	0.03	0.56	117
Africa	Africa	Malawi	Senegal	488	46,767	4.6	0	0.03	0.59	119
Africa	Africa	Malawi	The Gambia	414	20,349	2.55	0	0.02	0.45	94
Africa	Africa	Mali	Senegal	215	11,004	12.5	0	0.12	0.67	133
Africa	Africa	Mali	The Gambia	141	4,788	15.35	0	0.17	0.61	121

Africa	Africa	Senegal	The Gambia	188	7,467	22.24	0	0.24	0.77	133
Africa	Southeast Asia	DR of the Congo	Bangladesh	149	4,680	2.2	0	0.03	0.49	103
Africa	Southeast Asia	DR of the Congo	Cambodia	625	54,184	1.7	0	0.03	0.39	81
Africa	Southeast Asia	DR of the Congo	Laos	188	8,736	2.66	0	0.05	0.41	86
Africa	Southeast Asia	DR of the Congo	Myanmar	161	5,928	2.34	0	0.04	0.47	98
Africa	Southeast Asia	DR of the Congo	Thailand	244	14,560	1.41	0	0.03	0.48	102
Africa	Southeast Asia	DR of the Congo	Vietnam	200	9,984	2.18	0	0.04	0.46	97
Africa	Southeast Asia	Ghana	Bangladesh	608	25,335	1.75	0	0.02	0.49	103
Africa	Southeast Asia	Ghana	Cambodia	1084	29,3323	0.92	0	0.01	0.65	1285
Africa	Southeast Asia	Ghana	Laos	647	47,292	0.84	0	0.01	0.88	164
Africa	Southeast Asia	Ghana	Myanmar	620	32,091	2.44	0	0.03	0.61	124
Africa	Southeast Asia	Ghana	Thailand	703	78,820	1.7	0	0.02	0.56	113
Africa	Southeast Asia	Ghana	Vietnam	659	54,048	0.97	0	0.02	1.41	237
Africa	Southeast Asia	Guinea	Bangladesh	145	4,500	2.02	0	0.03	0.58	120
Africa	Southeast Asia	Guinea	Cambodia	621	52,100	4.07	0	0.05	0.41	86
Africa	Southeast Asia	Guinea	Laos	184	8,400	2.42	0	0.04	0.44	91
Africa	Southeast Asia	Guinea	Myanmar	157	5,700	4.98	0	0.07	0.55	112
Africa	Southeast Asia	Guinea	Thailand	240	14,000	3.3	0	0.05	0.56	116
Africa	Southeast Asia	Guinea	Vietnam	196	9,600	3.08	0	0.05	0.42	89
Africa	Southeast Asia	Malawi	Bangladesh	402	16,065	1.09	0	0.01	0.46	98
Africa	Southeast Asia	Malawi	Cambodia	878	185,997	0.71	0	0.01	0.45	94

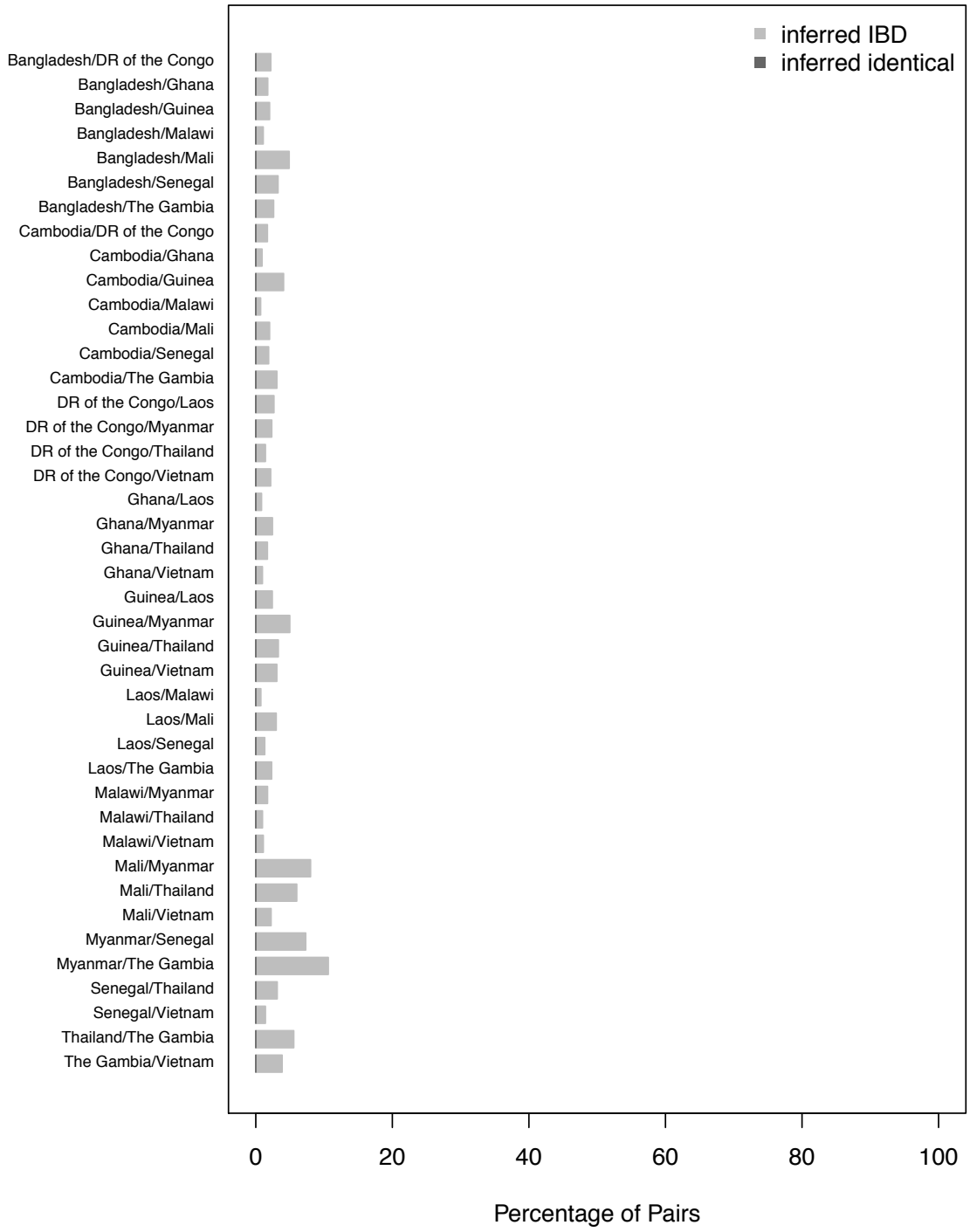
Africa	Southeast Asia	Malawi	Laos	441	29,988	0.74	0	0.01	0.45	93
Africa	Southeast Asia	Malawi	Myanmar	414	20,349	1.72	0	0.02	0.49	103
Africa	Southeast Asia	Malawi	Thailand	497	49,980	0.98	0	0.01	0.44	93
Africa	Southeast Asia	Malawi	Vietnam	453	34,272	1.1	0	0.01	0.51	107
Africa	Southeast Asia	Mali	Bangladesh	129	3,780	4.89	0	0.06	0.59	124
Africa	Southeast Asia	Mali	Cambodia	605	43,764	2.02	0	0.04	0.57	120
Africa	Southeast Asia	Mali	Laos	168	7,056	2.99	0	0.06	0.56	114
Africa	Southeast Asia	Mali	Myanmar	141	4,788	8.02	0	0.14	0.69	144
Africa	Southeast Asia	Mali	Thailand	224	11,760	6	0	0.11	0.58	120
Africa	Southeast Asia	Mali	Vietnam	180	8,064	2.24	0	0.04	0.53	108
Africa	Southeast Asia	Senegal	Bangladesh	176	5,895	3.26	0	0.03	0.55	115
Africa	Southeast Asia	Senegal	Cambodia	652	68,251	1.88	0	0.03	0.47	99
Africa	Southeast Asia	Senegal	Laos	215	11,004	1.31	0	0.03	0.65	138
Africa	Southeast Asia	Senegal	Myanmar	188	7,467	7.31	0	0.11	0.65	133
Africa	Southeast Asia	Senegal	Thailand	271	18,340	3.14	0	0.06	0.55	107
Africa	Southeast Asia	Senegal	Vietnam	227	12,576	1.41	0	0.03	0.69	145
Africa	Southeast Asia	The Gambia	Bangladesh	102	2,565	2.61	0	0.03	0.46	97
Africa	Southeast Asia	The Gambia	Cambodia	578	29,697	3.09	0	0.05	0.39	82
Africa	Southeast Asia	The Gambia	Laos	141	4,788	2.32	0	0.04	0.52	109
Africa	Southeast Asia	The Gambia	Myanmar	114	3,249	10.62	0	0.16	0.48	97
Africa	Southeast Asia	The Gambia	Thailand	197	7,980	5.56	0	0.09	0.47	99

Africa	Southeast Asia	The Gambia	Vietnam	153	5,472	3.86	0	0.06	0.48	102
Africa	Oceania	DR of the Congo	PNG	142	3,952	0.66	0	0.01	0.87	182
Africa	Oceania	Ghana	PNG	600	20,831	0.48	0	0.01	0.96	189
Africa	Oceania	Guinea	PNG	134	3,400	0.88	0	0.01	0.83	176
Africa	Oceania	Malawi	PNG	391	12,138	0.33	0	0	0.66	140
Africa	Oceania	Mali	PNG	122	3,192	0.5	0	0.01	3.57	749
Africa	Oceania	Senegal	PNG	169	4,978	0.26	0	0	1.38	288
Africa	Oceania	The Gambia	PNG	94	2,109	1.09	0	0.02	0.86	181
Southeast Asia	Southeast Asia	Bangladesh	Cambodia	566	23,445	4.08	0	0.04	0.46	94
Southeast Asia	Southeast Asia	Bangladesh	Laos	129	3,780	3.7	0	0.04	0.49	102
Southeast Asia	Southeast Asia	Bangladesh	Myanmar	102	2,565	7.99	0	0.07	0.52	105
Southeast Asia	Southeast Asia	Bangladesh	Thailand	185	6,300	6.19	0	0.05	0.51	106
Southeast Asia	Southeast Asia	Bangladesh	Vietnam	141	4,320	5	0	0.04	0.44	92
Southeast Asia	Southeast Asia	Cambodia	Laos	605	43,764	8.95	0	0.15	1.41	212
Southeast Asia	Southeast Asia	Cambodia	Myanmar	578	29,697	15.73	0	0.17	0.5	94
Southeast Asia	Southeast Asia	Cambodia	Thailand	661	72,940	23.08	0.2	1	3.34	294
Southeast Asia	Southeast Asia	Cambodia	Vietnam	617	50,016	19.22	0.06	0.39	1.4	188
Southeast Asia	Southeast Asia	Laos	Myanmar	141	4,788	10.07	0	0.11	0.45	89
Southeast Asia	Southeast Asia	Laos	Thailand	224	11,760	8.22	0	0.14	1.25	213
Southeast Asia	Southeast Asia	Laos	Vietnam	180	8,064	12.85	0	0.18	1.02	152
Southeast Asia	Southeast Asia	Myanmar	Thailand	197	7,980	41.37	0.01	0.55	0.94	134
Southeast Asia	Southeast Asia	Myanmar	Vietnam	153	5,472	17.05	0	0.16	0.54	104

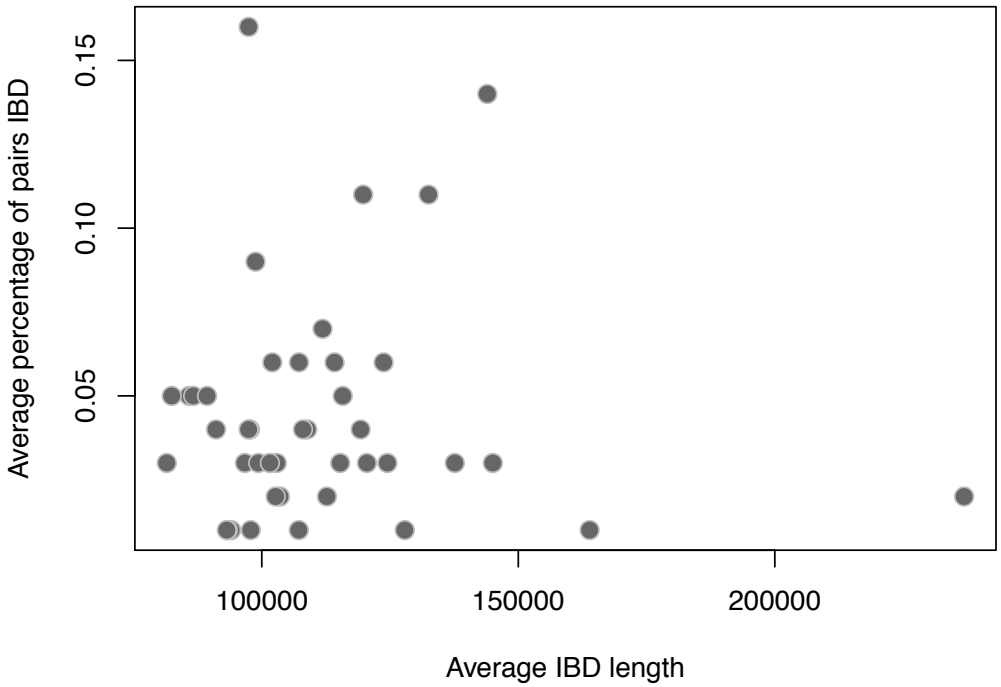
Southeast Asia	Southeast Asia	Thailand	Vietnam	236	13,440	20.52	0	0.2	0.75	126
Southeast Asia	Oceania	Bangladesh	PNG	83	1,710	0.53	0	0.01	0.53	112
Southeast Asia	Oceania	Cambodia	PNG	558	19,277	0.48	0	0.01	0.44	92
Southeast Asia	Oceania	Laos	PNG	121	3,108	0.13	0	0	1.26	266
Southeast Asia	Oceania	Myanmar	PNG	94	2,109	0.14	0	0	0.63	133
Southeast Asia	Oceania	Thailand	PNG	177	5,180	0.29	0	0	0.69	147
Southeast Asia	Oceania	Vietnam	PNG	133	3,552	0.48	0	0	0.29	62



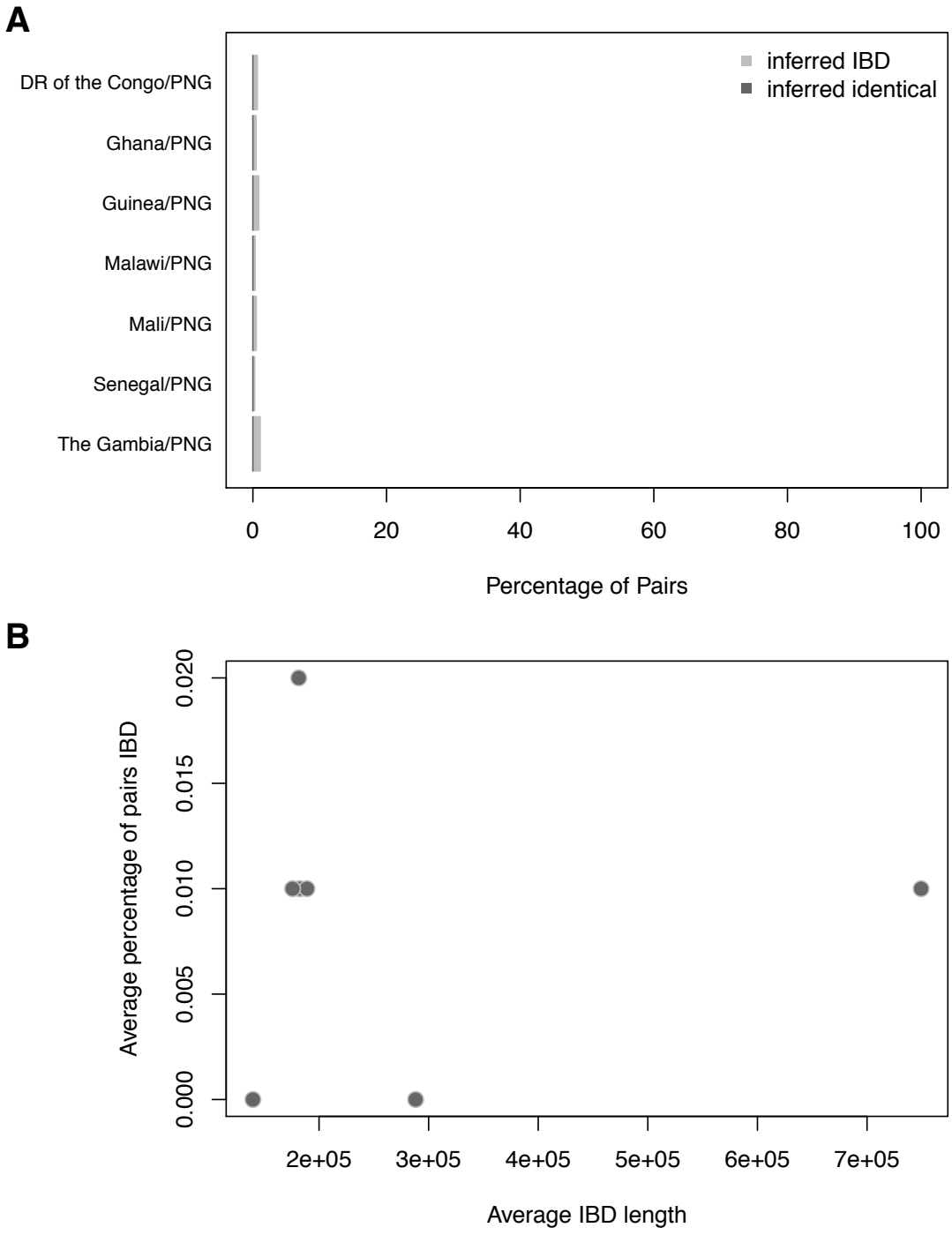
Appendix C Figure 11. Summary of Appendix C Table 9. **A.** The percentage of pairs with any inferred IBD at least 50 kb in length between countries in Africa and the percentage of pairs with identical genomes. **B.** The average percentage of pairs IBD against the average IBD length in base-pairs.



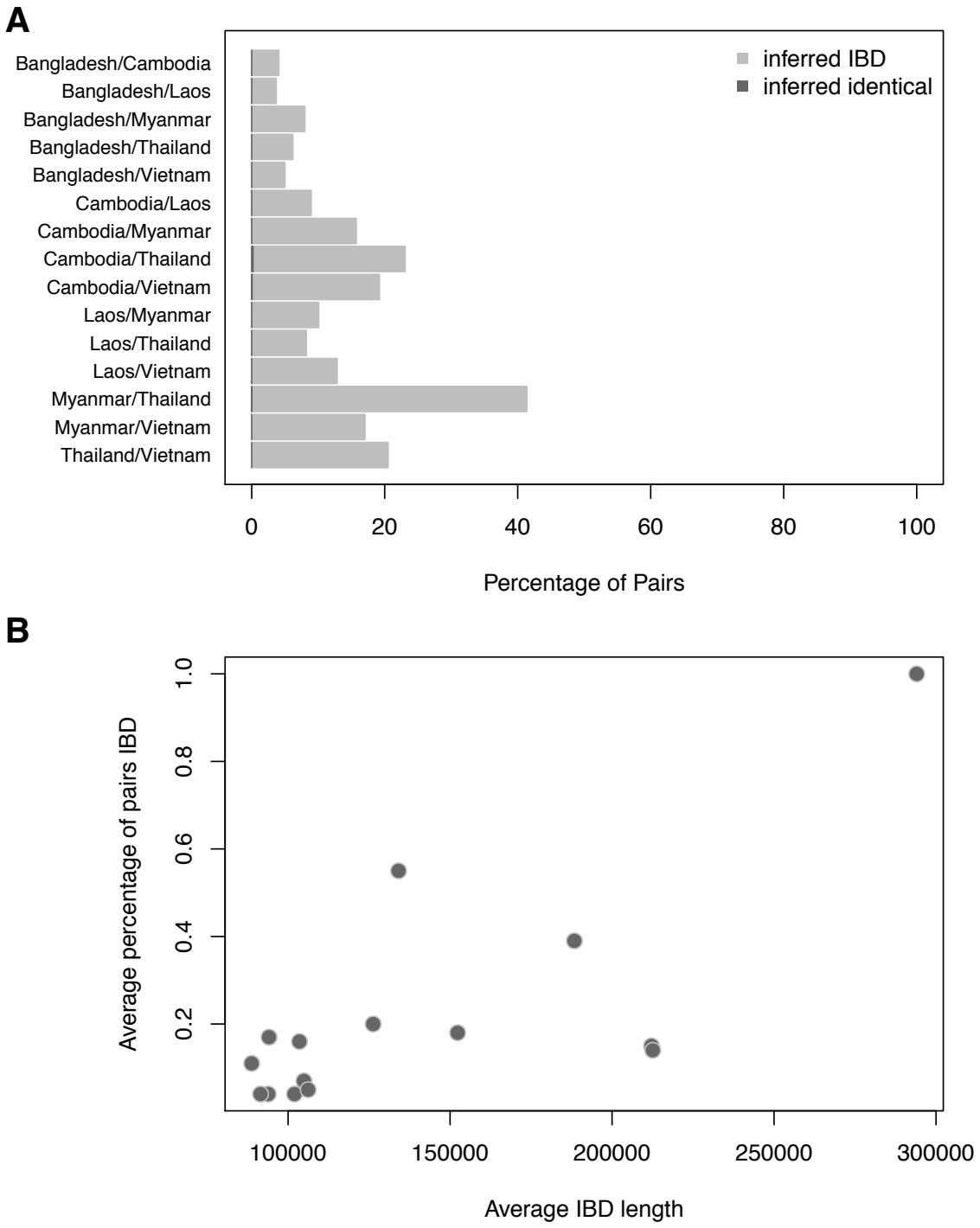
Appendix C Figure 12. Summary of Appendix C Table 9. The percentage of pairs with any inferred IBD at least 50 kb in length between countries in Africa and Southeast Asia and the percentage of pairs with identical genomes.



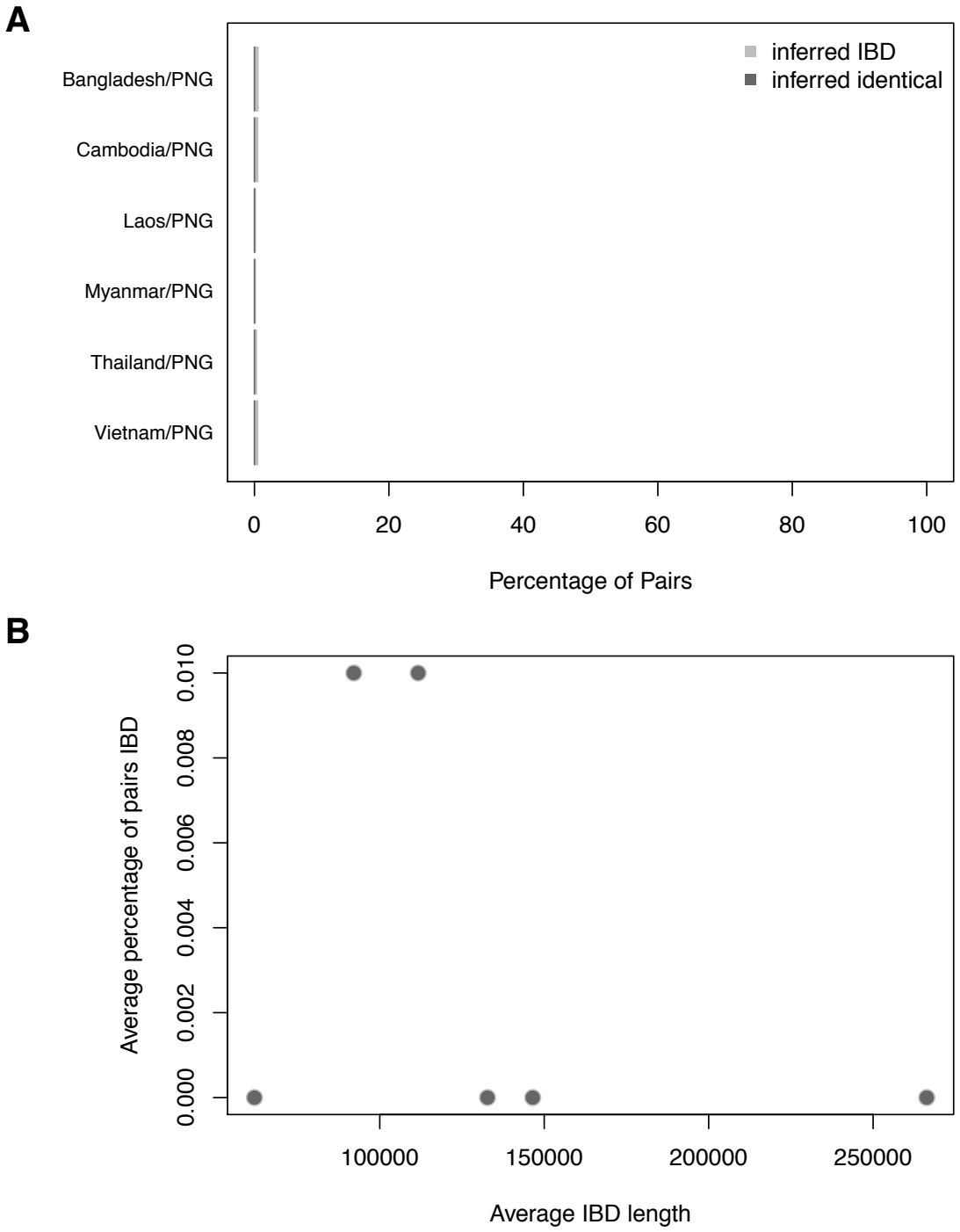
Appendix C Figure 13. Summary of Appendix C Table 9. The average percentage of pairs IBD against the average IBD length (bp) between countries in Africa and Southeast Asia.



Appendix C Figure 14. Summary of Appendix C Table 9. **A.** The percentage of pairs with any inferred IBD at least 50 kb in length between Africa and PNG and the percentage of pairs with identical genomes. **B.** The average percentage of pairs IBD against the average IBD length in base-pairs.



Appendix C Figure 15. Summary of Appendix C Table 9. **A.** The percentage of pairs with any inferred IBD at least 50 kb in length between Southeast Asian countries and the percentage of pairs with identical genomes. **B.** The average percentage of pairs IBD against the average IBD length in base-pairs.



Appendix C Figure 16. Summary of Appendix C Table 9. **A.** The percentage of pairs with any inferred IBD at least 50 kb in length between Southeast Asia and PNG and the percentage of pairs with identical genomes. **B.** The average percentage of pairs IBD against the average IBD length in base-pairs.

Appendix C Table 10. The top 5 selection signals within each country and the IBD proportion/X_{iR} test statistic for the SNP with the largest $-\log_{10}$(p-value) within the selection interval.

Country	Chromosome	Start	End	IBD percentage	X_{iR}
DR of the Congo	6	1040011	1292764	1.34	315.48
DR of the Congo	7	383527	688642	0.37	40.67
DR of the Congo	10	1421301	1570694	0.24	31.95
DR of the Congo	11	1950959	2003228	0.26	10.67
DR of the Congo	12	761260	909881	0.5	39.51
Ghana	2	756399	862030	0.27	39.44
Ghana	6	1070251	1294804	0.84	70.96
Ghana	7	210646	699826	0.13	77.21
Ghana	10	1350374	1571407	0.41	28.63
Ghana	12	721251	968375	0.53	69.44
Guinea	4	614998	698534	1.03	48.44
Guinea	6	1050336	1294804	1.39	269.69
Guinea	7	350609	659941	0.99	79.74
Guinea	12	760468	1029068	1.52	92.96
Guinea	12	1745700	1897010	0.67	30.13
Malawi	6	1081929	1294804	1	116.74
Malawi	8	372767	779813	0.69	150.8
Malawi	10	1400002	1570653	0.36	79.86
Malawi	11	1790115	2003228	0.45	18.07
Malawi	12	780748	1079903	0.71	68.74
Mali	6	1001323	1292769	3.82	79.33
Mali	7	360243	679897	3.24	109.95
Mali	10	1430096	1571407	1.03	46.47
Senegal	2	720032	861423	0.72	4.68
Senegal	3	901428	993530	0.72	5.45
Senegal	4	615405	748410	2.14	9.74
Senegal	6	1050336	1292769	6.62	323.28
Senegal	7	252368	619957	3.01	74.82
The Gambia	4	540033	779909	7.64	1447.64
The Gambia	4	930011	1143943	2.07	19.4
The Gambia	7	331248	719023	2.51	50.01
The Gambia	8	105218	219366	2.07	12.62
The Gambia	8	420318	649599	2.57	27.68
Bangladesh	4	620279	679065	0.71	19.87
Bangladesh	6	1191624	1294070	1.01	95.06
Bangladesh	7	175076	718355	1.62	69.07

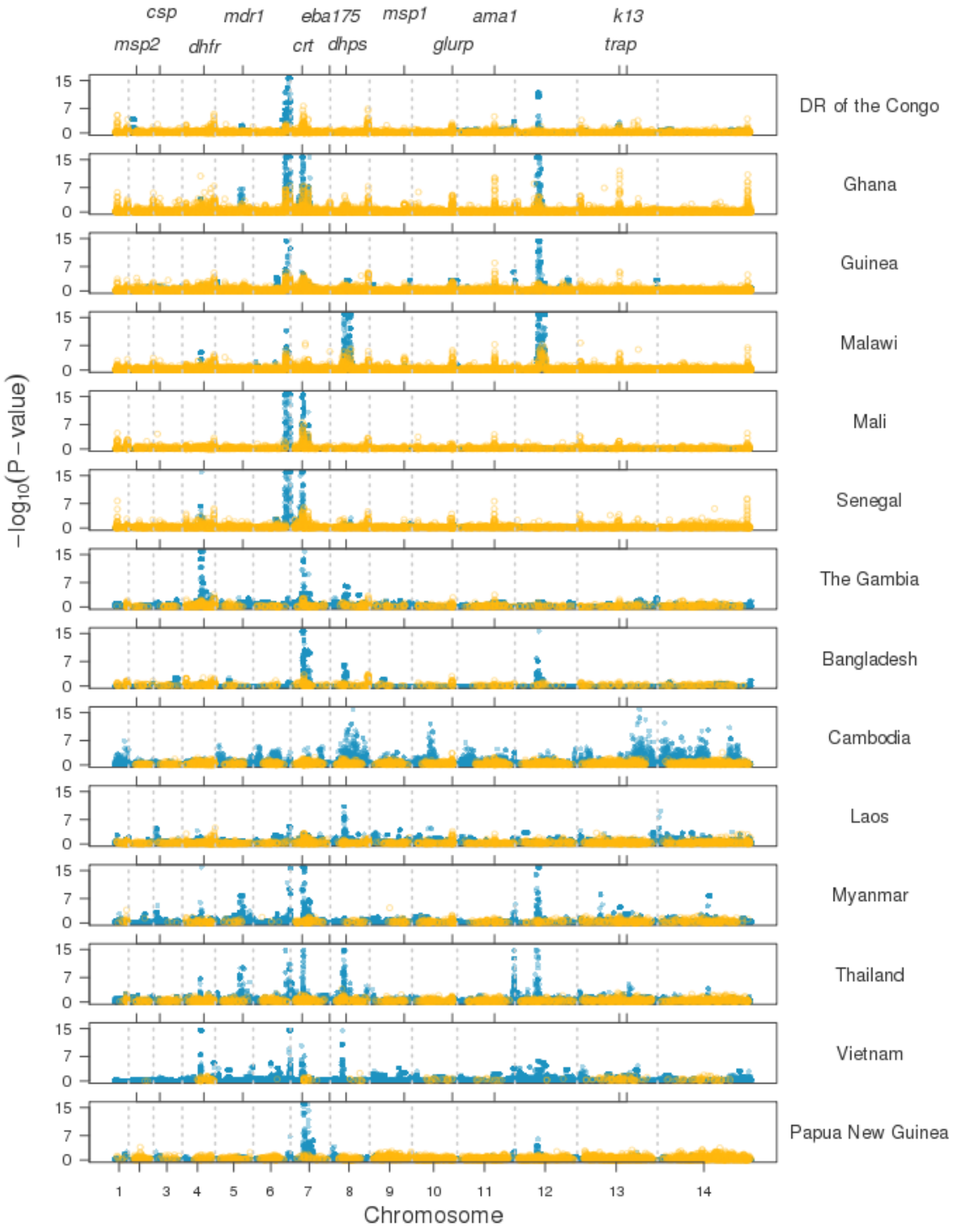
Bangladesh	8	467370	580545	2.02	695.78
Bangladesh	12	766623	989679	0.91	99.4
Cambodia	8	420692	979511	8.94	64.92
Cambodia	10	550118	799419	6.8	38.57
Cambodia	13	1901748	2389910	9.26	73.76
Cambodia	14	35796	999557	4.92	46.27
Cambodia	14	1250576	1999655	6.8	65.82
Laos	1	93378	199493	2.81	25.87
Laos	3	71014	359186	3.18	68.25
Laos	4	540009	899970	3.82	201.13
Laos	8	425234	599944	3.56	55.12
Laos	13	75051	139950	3.18	71.22
Myanmar	4	545154	697350	7.52	76.87
Myanmar	6	1020329	1294426	8.15	110.64
Myanmar	7	450016	739678	4.82	36.25
Myanmar	11	1890328	2003228	4.26	50.42
Myanmar	12	631425	895415	10.34	94.4
Thailand	6	1002288	1294426	6.1	72.33
Thailand	7	405600	679897	7.87	80.05
Thailand	8	415327	698268	13.71	80.57
Thailand	11	1910597	2003233	6.55	81.83
Thailand	12	701638	977430	9.48	90.67
Vietnam	4	540009	679043	6.07	45.38
Vietnam	6	742857	959371	5.26	37.65
Vietnam	9	79451	498806	5.26	38.38
Vietnam	11	1900435	2003233	6.1	100.19
Vietnam	12	561246	1246071	6.71	113.73
PNG	5	803678	1099386	1.8	10.89
PNG	7	357431	849589	6.76	131.64
PNG	11	1900435	2001345	2.1	9.71
PNG	12	750192	999517	2.25	25.58
PNG	12	1680129	2162665	2.4	17.74

Appendix C Table 11. The number of MOI = 1 isolates and SNPs included in the analysis of selection signatures within countries using iHS and isoRelate.

Region	Country	No. isolates	No. SNPs
Africa	DR of the Congo	47	34,382
Africa	Ghana	243	30,361
Africa	Guinea	51	61,546
Africa	Malawi	134	57,302
Africa	Mali	40	26,316
Africa	Senegal	110	21,317
Africa	The Gambia	40	35,059
Southeast Asia	Bangladesh	21	20,928
Southeast Asia	Cambodia	396	28,101
Southeast Asia	Laos	49	35,391
Southeast Asia	Myanmar	45	26,843
Southeast Asia	Thailand	107	25,335
Southeast Asia	Vietnam	69	32,761
Oceania	PNG	30	16,164

Appendix C Table 12. Positive control genes in *P.falciparum* where there is evidence of selection in either drug resistance, vaccine candidate or anti-folate resistance genes. The gene identifiers, names and locations were obtained from PlasmoDB. The type of signal (balancing or positive) and corresponding references are also given.

Gene ID	Name	Location	Description	Signal	Reference
PF3D7_1133400	AMA1	Pf3D7_11_v3: 1,293,856 - 1,295,724 (+)	apical membrane antigen 1	balancing	Polley, S. D. & Conway, D. J. (2001)
PF3D7_0709000	CRT	Pf3D7_07_v3: 403,222 - 406,317 (+)	chloroquine resistance transporter	positive	Mu, J. *et al*. (2010)
PF3D7_0304600	CSP	Pf3D7_03_v3: 221,323 - 222,516 (-)	circumsporozoite protein	balancing	Weedall, G. D., et al. (2007)
PF3D7_0417200	DHFR	Pf3D7_04_v3: 748,088 - 749,914 (+)	bifunctional dihydrofolate reductase-thymidylate synthase	positive	Nwakanma, D. C. *et al*. (2014)
PF3D7_0810800	DHPS	Pf3D7_08_v3: 548,200 - 550,616 (+)	dihydropteroate synthetase	positive	Nwakanma, D. C. *et al*. (2014)
PF3D7_0731500	EBA175	Pf3D7_07_v3: 1,358,055 - 1,362,929 (+)	erythrocyte binding antigen-175	balancing	Baum, J et al. (2003)
PF3D7_1035300	GLURP	Pf3D7_10_v3: 1,399,195 - 1,402,896 (+)	glutamate-rich protein	balancing	Conway, D.J. (1997)
PF3D7_1343700	K13	Pf3D7_13_v3: 1,724,817 - 1,726,997 (-)	kelch protein K13	positive	Miotto, O. *et al* (2013)
PF3D7_0523000	MDR1	Pf3D7_05_v3: 957,890 - 962,149 (+)	multidrug resistance protein	positive	Nwakanma, D. C. *et al*. (2014)
PF3D7_0930300	MSP1	Pf3D7_09_v3: 1,201,812 - 1,206,974 (+)	merozoite surface protein 1	balancing	Tetteh, K. K. A. *et al*. (2009)
PF3D7_0206800	MSP2	Pf3D7_02_v3: 273,689 - 274,507 (-)	merozoite surface protein 2	balancing	Tetteh, K. K. A. *et al*. (2009)
PF3D7_1335900	TRAP	Pf3D7_13_v3: 1,464,895 - 1,466,619 (-)	thrombospondin-related adhesion protein	positive	Weedall, G. D., et al. (2007)



Appendix C Figure 17. Comparison of $-\log_{10}$(p-values) generated by isoRelate (blue) and iHS (yellow) within each country. Grey dashed vertical lines indicate chromosome boundaries and ticks on the upper x-axis mark twelve positive control genes (Supplementary Table 10)